

ePCA: High Dimensional Exponential Family PCA

Lydia T. Liu*
Princeton University
ltliu@princeton.edu

Edgar Dobriban*
Stanford University
dobriban@stanford.edu

Amit Singer
Princeton University
amits@math.princeton.edu

November 18, 2016

Abstract

Many applications involve large collections of high-dimensional datapoints with noisy entries from exponential family distributions. It is of interest to estimate the covariance and principal components of the noiseless distribution. In photon-limited imaging (e.g. XFEL) we want to estimate the covariance of the pixel intensities of 2-D images, where the pixels are low-intensity Poisson variables. In genomics we want to estimate population structure from biallelic—Binomial(2)—genetic markers such as Single Nucleotide Polymorphisms (SNPs). A standard method for this is Principal Component Analysis (PCA). However, PCA loses some of its optimality properties for non-Gaussian distributions and can be inefficient when applied directly.

We develop *ePCA* (exponential family PCA), a methodology for PCA on exponential family distributions. *ePCA* can be used for dimensionality reduction and denoising of large data matrices. It involves the eigendecomposition of a new covariance matrix estimator, and is as fast as PCA. It is suitable for datasets with multiple types of variables.

The first step of *ePCA* is a diagonal debiasing of the sample covariance matrix. We obtain the convergence rate for covariance matrix estimation, and the Marchenko-Pastur law in high dimensions. Another key step of *ePCA* is *whitening*, a specific variable weighting. For SNPs, this recovers the widely used Hardy-Weinberg equilibrium (HWE) normalization. We show that whitening improves the signal strength, providing justification for HWE normalization. *ePCA* outperforms PCA in simulations as well as in XFEL and SNP data analysis. An open-source implementation is [available](#).

1 Introduction

In many applications we have large collections of high-dimensional data vectors with entries sampled from exponential families (such as Poisson or Binomial). This setting arises in image processing, computational biology, and natural language processing, for instance, in photon-limited imaging (e.g., [Luisier et al., 2011](#)) and in single-cell RNA-sequencing (e.g., [Stegle et al., 2015](#)). It is often of interest to reduce the dimensionality and understand the structure of the data.

The standard method for dimension reduction and denoising of large datasets is Principal Component Analysis (PCA) (e.g., [Jolliffe, 2002](#); [Anderson, 2003](#); [Muirhead, 2009](#)). However, PCA is most naturally designed for Gaussian data, and there is no commonly agreed upon extension to non-Gaussian settings such as exponential families (see, e.g., [Jolliffe, 2002](#), Sec. 14.4). While there are several proposals for extending PCA to non-Gaussian distributions, each of them has certain limitations, such as computational intractability for large datasets (see Sec. 2 for a detailed discussion).

We propose the new method *ePCA* for PCA of data from exponential families. *ePCA* involves the eigendecomposition of a new covariance matrix estimator. Like usual PCA, it can be used for visualization and denoising of large data matrices. Moreover, *ePCA* has several appealing properties. First, it is a flexible method suitable for datasets with multiple types of variables (such as Poisson, Binomial, and Negative

*The first two authors contributed equally to this work.

Binomial). Second, it is computationally efficient: it is as fast as usual PCA and scales to “big” datasets. Third, it has extensive theoretical justification, including for high dimensional data. We provide finite-sample convergence rates for the method, and a precise high-dimensional analysis building on random matrix theory. Fourth, each step of *ePCA* is interpretable, which can be important to practitioners.

We perform extensive simulations with *ePCA* and show that in several metrics it outperforms usual PCA and variants such as PCA after standardization. We apply *ePCA* to simulated X-ray Free Electron Laser (XFEL) data, where it leads to better denoising than PCA. We also apply *ePCA* to a dataset from the Human Genome Diversity Project (HGDP) measuring Single Nucleotide Polymorphisms, where it leads to a clearer structure in the principal component (PC) scores than PCA.

ePCA is publicly available in an open-source Matlab implementation from github.com/lydiatliu/epca/. That link also has software to reproduce our computational results.

To motivate our method, we now discuss a few potential application areas.

1.1 Denoising XFEL diffraction patterns

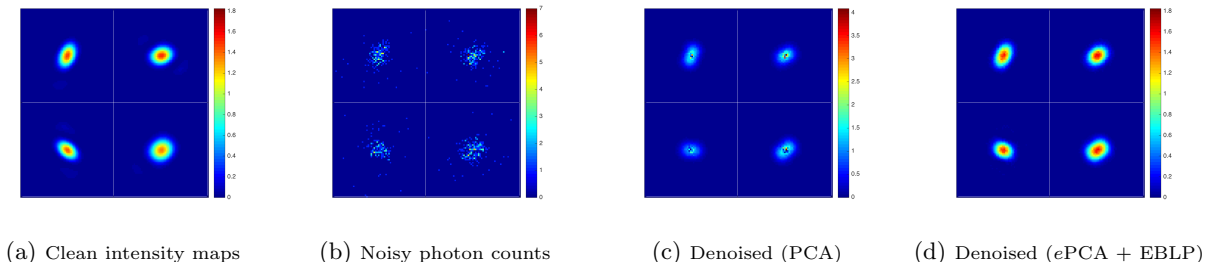


Figure 1.1: XFEL diffraction pattern formation model and denoising. See section 6.1 for details.

X-ray free electron lasers (XFEL) are an exciting experimental technique used to understand the three-dimensional structure of molecules (e.g., [Gaffney and Chapman, 2007](#)). XFEL is rapidly becoming more popular, see e.g. the recent special issues on XFEL, [Favre-Nicolin et al. \(2015\)](#); [Maia and Hajdu \(2016\)](#). Many datasets are publicly shared on platforms such as cxi.db.org. A key advantage of this imaging technique is that it uses extremely short femtosecond X-ray pulses, during which the molecule does not change its structure. However, images only contain information about the magnitude of the Fourier transform, leading to the additional phase retrieval problem.

XFEL imaging leads to two-dimensional diffraction patterns of single particles. As illustrated in Figure 1.1, these images are very noisy due to the low number of photons, and the count-noise at each detector follows an approximately Poisson distribution. Further, we only capture one diffraction pattern per particle, and the particle orientations are unknown.

In order to reconstruct the 3-D structure of the particle, one proposal is to estimate the unknown orientations (e.g., [Loh and Elser, 2009](#)). Alternatively, assuming that the orientations are uniformly distributed over the special orthogonal group $SO(3)$, Kam’s method from the related setting of cryo-electron microscopy (Cryo-EM) might provide a way to estimate the 3-D structure without estimating the orientations ([Kam, 1980](#)). Kam’s method for 3-D structure reconstruction requires estimation of the covariance matrix of the noiseless 2-D images. To properly work with the Poisson case, this motivates us to develop a method for covariance estimation in exponential families. To illustrate the improvement in covariance estimation of *ePCA* over PCA, in Figure 1.1 we show the result of denoising simulated XFEL using different estimated covariance matrices, where EBLP, or empirical best linear predictor, is the denoiser we develop in section 5 for use in conjunction with *ePCA*.

1.2 Single-cell RNA-sequencing data

Single-cell RNA-sequencing (scRNA-seq) is a new and rapidly developing experimental methodology in genomics that allows scientists to probe information at the individual cell level, to an unprecedented degree of granularity (see e.g., [Stegle et al., 2015](#), for a review). In scRNA-seq, we observe read counts of many genes extracted from a large number of individual cells. Suppose X_{ij} is the number of reads mapped to gene j for sample cell i . Then popular models assume that X_{ij} follows a negative binomial or a Poisson distribution (e.g., [Grün et al., 2014](#)). In the latter case, $X_{ij} \sim \text{Poisson}(\lambda_{ij})$ where λ_{ij} represents the rate at which reads map to gene j relative to other genes in cell i .

Reads from single cells typically have many zeros, and therefore the Poisson or negative binomial models are more appropriate than a Gaussian approximation. Dimension reduction of such datasets is important in exploratory analyses that seek to understand cell-to-cell heterogeneity or clustering. PCA is already used in standard pipelines (e.g., [Macosko et al., 2015](#)), however without explicitly taking into account non-Gaussianity. This serves as a motivation for our methods.

1.3 Genetic polymorphism data/SNPs

Single Nucleotide Polymorphism (SNP) data is a commonly available datatype in genomics, used in thousands of studies of common traits and diseases. SNPs are the basis of Genome-Wide Association Studies (GWAS), which have led to hundreds of novel associations between common traits and genetic variants (e.g., [Visscher et al., 2012](#)).

SNP data can be represented as an $n \times p$ matrix X with X_{ij} equal to the number of minor alleles (0, 1 or 2) of the j -th SNP in the genome of the i -th individual. The number of individuals n can be more than 10,000, while the number of SNPs can be as large as 2.5 million. Binomial models are natural for such data.

PCA is commonly used to infer population structure from SNP data, with a wide range of applications, including correcting for confounding in GWAS (see e.g., [Patterson et al. \(2006\)](#)). It is thus of interest to understand the proper way to estimate the covariance matrix and PCs.

1.4 Our contributions

We now briefly summarize our contributions:

1. We propose the new method ePCA for PCA of exponential family data. PCA is based on the eigenvalues and eigenvectors of the sample covariance matrix, whereas ePCA is based on a new covariance estimator that we propose. This estimator is developed through a sequence of steps, each leading to increasingly accurate covariance matrix estimators (Sec. 3 and 4).

Our model assumes that the data is an i.i.d. sample from an exponential family whose *mean parameter*—the “clean” signal—is a *random vector from an unknown low-dimensional space* (Sec. 3.1). Under this hierarchical model, we propose a *diagonal debiasing* of the sample covariance matrix (Sec 3.2), and characterize the finite-sample rate of convergence to the true covariance matrix of the clean signals (Sec. 3.2.1).

2. To improve performance in high-dimensional settings, we propose a method of *whitening*, *shrinkage*, and *recoloring* the debiased covariance matrix (Sec. 4). Whitening is a special form of variable weighting, different from the usual method of standardizing the features to have unit norm. We justify it by proving the standard Marchenko-Pastur law ([Marchenko and Pastur, 1967](#)) for the asymptotic spectrum of the whitened sample covariance matrix (Sec. 4.1.1), and by showing that whitening improves the signal strength (Sec. 4.2.2). We also note that for biallelic genetic markers such as Single Nucleotide Polymorphisms (SNPs), whitening agrees with the widely used normalization assuming Hardy-Weinberg equilibrium (HWE) (Sec. 4.3). This provides perhaps the first theoretical justification for HWE normalization.

3. e PCA includes an eigenvalue shrinkage step to improve performance in high dimensions (Sec. 4.2). However, it turns out that the well-understood shrinkage methods for homoskedastic Gaussian distributions (e.g., Donoho et al., 2013) are suboptimal, and better estimation can be achieved with more shrinkage. We propose a *scaled* covariance estimator to improve performance (Sec 4.2.3). e PCA consists of the eigendecomposition of this estimator.

We evaluate our covariance estimator in a simulation study, and show that it reduces the MSE for covariance, eigenvalue, and eigenvector estimation (Sec. 4.2.3).

4. We apply e PCA to develop a new denoising method. Our method is a special empirical Best Linear Predictor (EBLP) from random effects models (Searle et al., 2009, Sec. 7.4) (Sec. 5), where we use our covariance estimator to estimate parameters in the BLP denoiser. In other areas such as electrical engineering and signal processing, the BLP is known as the “Wiener filter”, the “Linear Minimum Mean Squared Estimator (LMMSE)” (Kay, 1993, Ch. 12), the “linear Wiener estimator” (Mallat, 2008, p. 538), or the “optimal linear filter” (MacKay, 2003, p. 550-551).
5. We apply e PCA denoising to simulated XFEL data where it leads to better denoising than PCA (Sec. 6.1) We also apply e PCA to a SNP dataset from the Human Genome Diversity Project (HGDP) (Li et al., 2008), where it leads to a clearer structure in the PC scores than PCA (Sec. 6.6).

2 Related work

To give context for our method, we review related work. The reader intersted in the methodology can skip directly to Section 3. We refer to Jolliffe (2002) for a detailed overview of PCA methodology, to Anderson (2003); Muirhead (2009) for a more general overview of multivariate statistical analysis including PCA, to Johnstone (2007); Paul and Aue (2014); Yao et al. (2015) for discussions of high-dimensional statistics, random matrix theory and PCA.

2.1 Standardization and weighting in PCA

In applying PCA, one of the key concerns is whether or not to standardize the variables. Jolliffe (2002), Sec. 2.3, provides a detailed discussion of the advantages and disadvantages of standardization. Briefly, standardization ensures that results for different sets of random variables are more comparable, and also that PCs are less dominated by individual variables with large variances, including due to unit choice. Statistical inference can sometimes be more convenient without standardization. In exploratory analyses, however, standardization is usually preferred. Our results show that, in the special case of exponential family spiked models, our whitening method (Sec. 4.1) has several advantages over standardization.

A more general class of methods is *weighted PCA*, where PCA is applied to rescaled random variables $w_j X(j)$, for some $w_j > 0$ (Jolliffe, 2002, Sec. 2.3., Sec. 14.2.) In general, choosing the weights can be a nontrivial task. Some of the suggestions beyond dividing by standard errors include dividing by the range or mean of random variables (Gower, 1966). Our whitening step of e PCA (Sec. 4.1) is a particular weighting method, which is justified for data from exponential families. In addition to proposing the method, we also provide several theoretical justifications for it: the standard Marchenko-Pastur law, and the improvements in SNR (see Sec. 4.1).

2.2 PCA in non-Gaussian distributions

There have been several approaches suggested for extending PCA to non-Gaussian distributions, (see. e.g., Jolliffe, 2002, Sec. 14.4). One possibility is to use robust estimates of the covariance matrix (see. Jolliffe, 2002, Sec. 14.4, for references). Another approach is to change the quadratic objective function maximized by PCA into one that is suitable for non-Gaussian distributions, such as predictive power (e.g., Qian et al., 1994).

For data from exponential families, [Collins et al. \(2001\)](#) proposed an extension of PCA assuming that the natural parameter lies in a low dimensional space. Their approach was to factorize the matrix of natural parameter values by maximizing the log-likelihood. This lead to a non-convex optimization problem for which they propose an alternating maximization method without global convergence guarantees. In our settings, it is often more reasonable to model the mean parameter of the exponential family—e.g., the clean images in image analysis—as having a low rank structure. For instance, in an XFEL experiment, the images can be thought of as “low-complexity” perturbations of the mean image according to the state of the molecule. For this reason, our methods are not directly comparable in theory or in simulations. Other likelihood-based methods include [Sajama and Orlitsky \(2004\)](#); [Li and Tao \(2010\)](#). In contrast, our approach avoids high-dimensional optimization problems and provides precise results about the performance of the method by building on random matrix theory.

2.3 Denoising and covariance estimation by singular value shrinkage

Recently, results from random matrix theory have been used for studying covariance estimation and PCA for Gaussian and rotationally invariant data (e.g., [Shabalin and Nobel, 2013](#); [Donoho et al., 2013](#); [Gavish and Donoho, 2014](#); [Nadakuditi, 2014](#)). While the qualitative insights they identify—e.g., the improvements due to eigenvalue shrinkage—are relevant to our setting, the specific results and methods do not apply directly.

The recent work of [Bigot et al. \(2016\)](#) develops a generalized Stein’s Unbiased Risk Estimation (SURE) approach for singular value shrinkage denoising of low-rank matrices in exponential families. However, their shrinkage formulas become numerically intractable for Frobenius norm beyond Gaussian errors, and they instead introduce a heuristic algorithm in their simulations. Moreover in their Poisson-noise simulations, they work in a regime where optimal shrinkage reduces to no shrinkage (e.g., their Fig. 5a), possibly because the signal is very strong. In contrast, we work in a setting of much weaker signal-to-noise ratio (SNR), where the necessary singular value shrinkage is substantial.

2.4 Image processing and denoising

There are many approaches to denoising in image and signal processing. The vast majority are designed for Gaussian noise, see for instance [Elad \(2010\)](#); [Starck et al. \(2010\)](#). Some of the key approaches exploit sparsity, either in a fixed basis or dictionary—such as Fourier or wavelet—or in a basis that is estimated (or learned) from the data. Our setting is quite different, because we have many very noisy samples—e.g., XFEL images—whereas the classical setting has one moderately noisy image. Most classical methods are not designed for sharing information across multiple images.

[Starck et al. \(2010\)](#) Sec. 6.5. provides an overview of the classical single-image denoising methods for Poisson noise. Popular approaches reduce to the Gaussian case by a wavelet transform such as a Haar transform, followed by Wiener filtering ([Nowak and Baraniuk, 1999](#)), by adaptive wavelet shrinkage, or by approximate variance stabilization such as the Anscombe transform (e.g., [Donoho, 1993](#)). The latter one is known to work well for Poisson signals with large parameters, due to the approximate normality of the Poisson. However, the normal approximation breaks down for the Poisson with a small parameter, such as photon-limited XFEL (see e.g., [Starck et al., 2010](#), Sec. 6.6).

Other well-known methods are based on singular value thresholding (SVT), e.g., [Østergaard et al. \(1996\)](#); [Worsley et al. \(2005\)](#). Works differ in how they incorporate prior knowledge about the noise distribution. For example, [Furnival et al. \(2016\)](#) performs SVT of the data matrix of image time-series, picking the regularization parameter to minimize the Poisson-Gaussian Unbiased Risk Estimator (PGURE). We instead whiten the data and propose a second-moment based denoising method.

Alternatively, [Cao and Xie \(2014\)](#) frames denoising as a regularized maximum likelihood problem and uses SVT to optimize an approximation of the Poisson likelihood. Our approach avoids nonconvex likelihood computations. Another approach applies the alternating minimization-based PCA extension proposed in [Collins et al. \(2001\)](#) with refinements to exploit self-similarity in natural images ([Salmon et al., 2014](#)).

3 Covariance estimation

ePCA is the eigendecomposition of a new covariance matrix estimator that we propose. To develop this estimator, we start with the sample covariance matrix and propose a sequence of improvements.

We will work with observations Y from the canonical one-parameter exponential family with density

$$p_\theta(y) = \exp[\theta y - A(\theta)] \quad (1)$$

with respect to a σ -finite measure ν on \mathbb{R} (see e.g., [Lehmann and Romano \(2005\)](#)). Here $\theta \in \mathbb{R}$ is the natural parameter of the family and $A(\theta) = \log \int \exp(\theta y) d\nu(y)$ is the log-partition function. We assume the distribution is well-defined for all θ in an open set. The mean and variance of Y can be expressed as $\mathbb{E}Y = A'(\theta)$ and $\text{Var}[Y] = A''(\theta)$, where we denote $g'(\theta) = dg(\theta)/d\theta$.

Well-known examples include the normal distribution $y \sim \mathcal{N}(m, \sigma^2)$, with fixed σ^2 assumed to be known. Here the carrier measure has density $\nu(dy) = (2\pi\sigma^2)^{-1/2} \exp(-y^2/(2\sigma^2))dy$ with respect to Lebesgue measure on \mathbb{R} , while $\theta = m/\sigma^2$ and $A(\theta) = \sigma^2\theta^2/2$. A second example is the Poisson distribution $y \sim \text{Poisson}(x)$. Here the carrier measure is the discrete measure with density $\nu(dy) = 1/y!$ with respect to the counting measure on the non-negative integers, while $\theta = \log(x)$ and $A(\theta) = \exp(\theta)$.

3.1 The observation model

We observe n i.i.d. noisy data vectors $Y_i \in \mathbb{R}^p$ drawn from an unknown distribution. Let Y be a random vector with the same distribution. In the XFEL application, Y is the noisy image, while in single-cell RNA sequencing, Y is a vector of read counts for each gene. We consider the following hierarchical model for Y . First, a latent vector—or hyperparameter— $\theta \in \mathbb{R}^p$ is drawn from a probability distribution D with mean μ_θ and covariance matrix Σ_θ . Conditional on θ , the coordinates of $Y = (Y(1), \dots, Y(p))^\top$ are drawn independently from an exponential family $Y(j) \sim p_{\theta(j)}(y)$ defined in (1). Formally, denoting by \sim the mean and the covariance of a random vector:

$$\begin{aligned} \theta &\sim (\mu_\theta, \Sigma_\theta) \\ Y(j)|\theta(j) &\sim p_{\theta(j)}(y), \quad Y = (Y(1), \dots, Y(p))^\top. \end{aligned}$$

Therefore, the mean of Y conditional on θ is

$$X := \mathbb{E}(Y|\theta) = (A'(\theta(1)), \dots, A'(\theta(p)))^\top = A'(\theta),$$

so that the noisy data vector Y can be expressed as $Y = A'(\theta) + \tilde{\varepsilon}$, with $\mathbb{E}(\tilde{\varepsilon}|\theta) = 0$, while the marginal mean of Y is $\mathbb{E}Y = \mathbb{E}A'(\theta)$. Thus Y can be thought of as a noisy realization of the clean vector $X = A'(\theta)$. However, the latent vector θ is also random and varies from sample to sample. In the XFEL application, $A'(\theta)$ are the noiseless images. In the RNA-seq applications, $A'(\theta)$ are the mean read counts of all genes, for a given cell.

It is important that we model the *mean* $A'(\theta)$ of the exponential family as our clean signal, as opposed to the *natural parameter* θ . In many applications, it is reasonable that the mean of our noisy signals (e.g., images) have a “low-complexity” structure, such as lying on a low-dimensional linear subspace. It is less clear what the corresponding low-complexity structure may be for the natural parameter. As mentioned in Sec. 2, this is a key fact distinguishing our approach from prior work like [Collins et al. \(2001\)](#).

We thus have $Y = A'(\theta) + \text{diag}[A''(\theta)]^{1/2}\varepsilon$, where the coordinates of ε are conditionally independent and standardized given θ . Therefore, the covariance of Y conditional on θ is

$$\text{Cov}[Y|\theta] = \text{diag}[A''(\theta(1)), \dots, A''(\theta(p))] = \text{diag}[A''(\theta)].$$

The marginal covariance of Y is given by the law of total covariance:

$$\text{Cov}[Y] = \text{Cov}[\mathbb{E}(Y|\theta)] + \mathbb{E}[\text{Cov}[Y|\theta]] = \text{Cov}[A'(\theta)] + \mathbb{E} \text{diag}[A''(\theta)]. \quad (2)$$

Examples include

- Normal: $Y \sim \mathcal{N}(X, \sigma^2 I_p)$, where $X \in \mathbb{R}^p$ is random, and σ^2 is known. In this case, we can represent $Y = X + \sigma\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I_p)$. The mean equation is $\mathbb{E}Y = \mathbb{E}X$, while the covariance equation is

$$\text{Cov}[Y] = \text{Cov}[X] + \sigma^2 I_p.$$

- Poisson: $Y \sim \text{Poisson}_p(X)$, where $X \in \mathbb{R}^p$ is random. We can write $Y = X + \text{diag}(X)^{1/2}\varepsilon$. The natural parameter is the vector θ with $\theta(j) = \log X(j)$. Since $A'(\theta(j)) = A''(\theta(j)) = \exp(\theta(j)) = X(j)$, we see $\mathbb{E}Y = \mathbb{E}X$, and

$$\text{Cov}[Y] = \text{Cov}[X] + \mathbb{E} \text{diag}[X].$$

3.2 Diagonal debiasing

Table 1: Covariance estimators

Notation	Name	Formula	Defined in	Motivation
S	Sample covariance	$S = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^\top$	(3)	-
S_d	Diagonal debiasing	$S_d = S - \text{diag}[V(\bar{Y})]$	(4)	Hierarchy
S_w	Whitening	$S_w = D_n^{-1/2} S_d D_n^{-1/2}$	(5)	Heteroskedasticity
$S_{w,\eta}$	Shrinkage	$S_{w,\eta} = \eta(S_w)$	(6)	High dimensionality
S_r	Recoloring	$S_r = D_n^{1/2} S_{w,\eta} D_n^{1/2}$	(7)	Heteroskedasticity
S_s	Scaling	$S_s = \sum \hat{\alpha}_i \hat{v}_i \hat{v}_i^\top$, where $S_r = \sum \hat{v}_i \hat{v}_i^\top$	(12)	Heteroskedasticity

We will propose several estimators of increasing sophistication to estimate the covariance matrix $\Sigma_x = \text{Cov}[A'(\theta)]$ of the noiseless vectors $X_i = A'(\theta_i)$ (see Table 1). Clearly, due to the covariance equation (2), the sample covariance matrix of Y_i is biased for estimating the diagonal elements of Σ_x . Fortunately, this bias can be corrected. Indeed, we must estimate the noise variances $\mathbb{E}A''(\theta(j))$. We know that $\mathbb{E}Y(j) = \mathbb{E}A'(\theta(j))$, so it is natural to define estimators via the *variance map* of the exponential family, which takes a mean parameter $A'(\theta)$ into the associated variance parameter $A''(\theta)$. Formally,

$$V(m) = A''[(A')^{-1}(m)].$$

If the distribution of Y is non-degenerate, $A''(\theta) = \text{Var}_\theta(Y) > 0$, so A' is increasing and invertible, and the variance map is well-defined.

We define the sample covariance estimator

$$S = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^\top, \quad (3)$$

where $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$ is the sample mean. We estimate $\mathbb{E}A''(\theta)$ by $V(\bar{Y})$, and define the *diagonally debiased* covariance estimator

$$S_d = S - \text{diag}[V(\bar{Y})]. \quad (4)$$

Continuing with our examples, we have

- Normal: Since $A''(\theta) = \sigma^2$, we have $V(m) = \sigma^2$, and $S_d = S - \sigma^2 I_p$. As mentioned above, σ^2 is assumed to be known. When it is unknown, it is often estimable on a case-by-case basis such as from the corners of microscopy images (e.g., [Katsevich et al., 2015](#)).
- Poisson: Here $A'(\theta) = A''(\theta) = \exp(\theta)$, so $V(m) = m$, and $S_d = S - \text{diag}[\bar{Y}]$.

In these two examples the estimator is unbiased, because V is affine (e.g., constant in the first case, and linear in the second case). When V is non-linear, the estimator can become slightly biased.

3.2.1 The rate of convergence

Our first theoretical result characterizes the finite-sample rate of convergence of the diagonally debiased covariance estimator S_d , for any fixed n, p . This estimator is not a standard sample covariance matrix—which is inconsistent in our case when $n \rightarrow \infty$ and p is fixed—therefore it is necessary to study its rate of convergence from first principles.

For this we need to make a few technical assumptions. First, we assume that the mean-variance map V is Lipschitz with constant L . It is easy to check that this is true in our running examples, namely for the Gaussian and Poisson distributions. We also assume that the coordinates of the random vector θ are almost surely bounded, $\|\theta\|_\infty \leq B$. Since A' is continuous and invertible, this is equivalent to the boundedness of $A'(\theta)$. This is reasonable in the areas that we are interested in—XFEL imaging does not have infinite energy so we have an upper bound on the intensity of pixels. For single-cell RNA sequencing, there is also a physical limit on the expression rate of any gene in a cell due to energy and resource constraints. Finally we assume that $m_4 = \max_i \mathbb{E}[Y(i)^4] \geq C$ for some universal constant $C > 0$. This is reasonable, as it states that at least some entries of the random vector have non-vanishing magnitude.

Let \lesssim denote inequality up to constants not depending on n and p . Let $\|\cdot\|_{\text{Fr}}$ be the Frobenius norm and $\|\cdot\|$ be the operator norm. Our result, proved in Sec. A.1, is

Theorem 3.1 (Rate of convergence of debiased covariance estimator). The diagonally debiased covariance estimator S_d has the following rates of convergence. In Frobenius norm, with $\mu := \mathbb{E}Y = \mathbb{E}X = \mathbb{E}A'(\theta)$:

$$\mathbb{E}[\|S_d - \Sigma_x\|_{\text{Fr}}] \lesssim \sqrt{\frac{p}{n}} [\sqrt{p} \cdot m_4 + \|\mu\|].$$

In operator norm, with the dimensional constant $C(p) = 4(1 + 2\lceil \log p \rceil)$:

$$\mathbb{E}[\|S_d - \Sigma_x\|] \lesssim \sqrt{C(p)} \frac{(\mathbb{E}\|Y\|^4)^{1/2} + (\log n)^3 (\log p)^2}{\sqrt{n}} + \sqrt{\frac{p}{n}} \left[1 + \sqrt{\frac{p}{n}} + \|\mu\| \right].$$

The two error rates are both of interest, and complement each other. The error rate in Frobenius norm captures the deviation across all entries of the covariance matrix. Since the squared Frobenius norm is approximately a sum of p^2 squared deviations of sample means from population means, each of which is $O(n^{-1})$, we expect a rate of $p/n^{1/2}$ in Frobenius norm. If the bound m_4 on the fourth moment is of order one, that is exactly what we obtain. It is not completely clear that our rates are optimal, but this suggests so for the Frobenius norm.

In contrast the error rate in operator norm captures the deviation in the extreme eigenvalues of the error matrix. Our rate in operator norm is typically faster than the rate in Frobenius norm. For instance, in the XFEL application, it is reasonable to assume that the total intensity across all detectors is fixed as the resolution increases. This would lead to a fixed value for $\mathbb{E}\|Y\|^4$ that does not grow with n . Hence, in such a setting the error rate in operator norm can be as fast as $(p/n)^{1/2}$ while that in Frobenius norm is $p/n^{1/2}$.

Our proof of Thm. 3.1 exploits that exponential family random variables are sub-exponential, so we can use corresponding moment bounds. We also rely on operator-norm bounds for random matrices from Tropp (2016) and on moment bounds from Boucheron et al. (2005).

4 Whitening and shrinkage

4.1 Whitening

In the previous sections, we showed that the diagonally debiased sample covariance matrix converges at a rate $O(pn^{-1/2})$. Next we propose a shrinkage method to improve this estimator in the high dimensional regime where $n, p \rightarrow \infty$ and $p/n \rightarrow \gamma > 0$. As a preliminary step, it is helpful to whiten the empirical covariance matrix and remove the effects of heteroskedasticity. This allows us to get closer to the *standard spiked*

model (Johnstone, 2001) where the noise has the same variance for all features. In that setting covariance estimation via eigenvalue shrinkage has been thoroughly studied (Donoho et al., 2013).

The vector of noise variances affecting the different components is $\mathbb{E}[A''(\theta)]$. For a given signal $Y = A'(\theta) + \text{diag}[A''(\theta)]^{1/2}\varepsilon$, whitening transforms it to $Y_w = \text{diag}[A''(\theta)]^{-1/2}A'(\theta) + \varepsilon$. Since the diagonal correction $D_n = \text{diag}[V(\bar{Y})]$ estimates $\mathbb{E} \text{diag}[A''(\theta)]$, we define the *whitened* covariance estimator by

$$S_w = D_n^{-1/2}S_d D_n^{-1/2} = D_n^{-1/2}S D_n^{-1/2} - I_p. \quad (5)$$

Examples:

- For normal observations, whitening reduces to $S_w = S/\sigma^2 - I_p$.
- For Poisson observations, every entry of the noisy vector has to be divided by square root of the corresponding entry of the sample mean, so $S_w = \text{diag}[\bar{Y}]^{-1/2}S \text{diag}[\bar{Y}]^{-1/2} - I_p$.

Whitening is different from *standardization*, the classical method for removing heteroskedasticity. To standardize, each feature—e.g., pixel—is divided by its empirical standard deviation (e.g., Jolliffe, 2002, Sec. 2.3.). This ensures that all features have the same norm. The sample covariance matrix becomes a sample correlation matrix. In our case it turns out that this procedure “over-corrects”. The overall variance $\text{Var}[Y(i)]$ of each feature is the sum of the signal variance $\text{Var}[A'(\theta(i))]$ and the noise variance $\mathbb{E}[A''(\theta(i))]$. Whitening divides by the estimated noise standard errors, while standardization divides by the *overall* standard error due to the signal and noise. Therefore, in our setting whitening is more justified than standardization. This is explained in more detail in a simulation in Sec. A.3.

4.1.1 Marchenko-Pastur law

A key advantage of whitening is that the whitened estimator has a simple well-understood asymptotic behavior. In contrast, the unwhitened estimator has a more complicated behavior. In this section, we show both of the above claims. We show that the limit spectra of our covariance matrix estimators are characterized by the Marchenko-Pastur (MP) law (Marchenko and Pastur, 1967), proving the general MP law for the sample covariance S , and the standard MP law for the whitened covariance S_w .

For simplicity, we consider the case is when $\theta \in \mathbb{R}^p$ is fixed. This can be thought of as the “null” case, where all mean signals are the same. Then we can write $Y_i = A'(\theta) + \text{diag}[A''(\theta)]^{1/2}\varepsilon_i$, where ε_i have independent standardized entries. Therefore, letting \mathcal{Y} be the $n \times p$ matrix whose rows are Y_i^\top , we have $\mathcal{Y} = \bar{1}A'(\theta)^\top + \mathcal{E} \text{diag}[A''(\theta)]^{1/2}$, where $\bar{1} = (1, 1, \dots, 1)^\top$ is the vector of all ones, and \mathcal{E} is an $n \times p$ matrix of independent standardized random variables.

Let now H_p denote the uniform distribution function on the p scalars $A''(\theta(i))$, $i = 1, \dots, p$. We assume that $A''(\theta(i)) > c$ for some universal constant $c > 0$. In the Poisson example, this means that the individual rates $x(i)$ are bounded away from 0. The reason for this assumption is to avoid the very sparse regime, where only a few nonzero entries per row are observed. In that case, the MP law is not expected to hold.

Consider the high dimensional asymptotic limit when $n, p \rightarrow \infty$ so that $p/n \rightarrow \gamma > 0$. Suppose moreover that H_p converges in distribution to some limit distribution, i.e., $H_p \Rightarrow H$. Since $\text{diag}[A''(\theta)]$ can be viewed as the population covariance matrix of the noise, H is the limit population spectral distribution (PSD). Since \mathcal{E} has independent standardized entries with bounded moments, it follows that the distribution of the p eigenvalues of $n^{-1}\mathcal{Y}^\top\mathcal{Y}$ converges almost surely to the general Marchenko-Pastur distribution $F_{\gamma,H}$ (Bai and Silverstein, 2009, Thm. 4.3).

Now, the sample covariance matrix S is a rank-one perturbation of $n^{-1}\mathcal{Y}^\top\mathcal{Y}$. Therefore its eigenvalue distribution also converges to the MP law. We state this for comparison with the next result.

Proposition 4.1 (Marchenko-Pastur law for sample covariance matrix). The eigenvalue distribution of S converges almost surely to the general Marchenko-Pastur distribution $F_{\gamma,H}$.

Since the general MP law has a complicated implicit description that needs to be studied numerically (see e.g., Dobriban, 2015), it is useful to work with the whitened covariance matrix S_w . Indeed, we establish

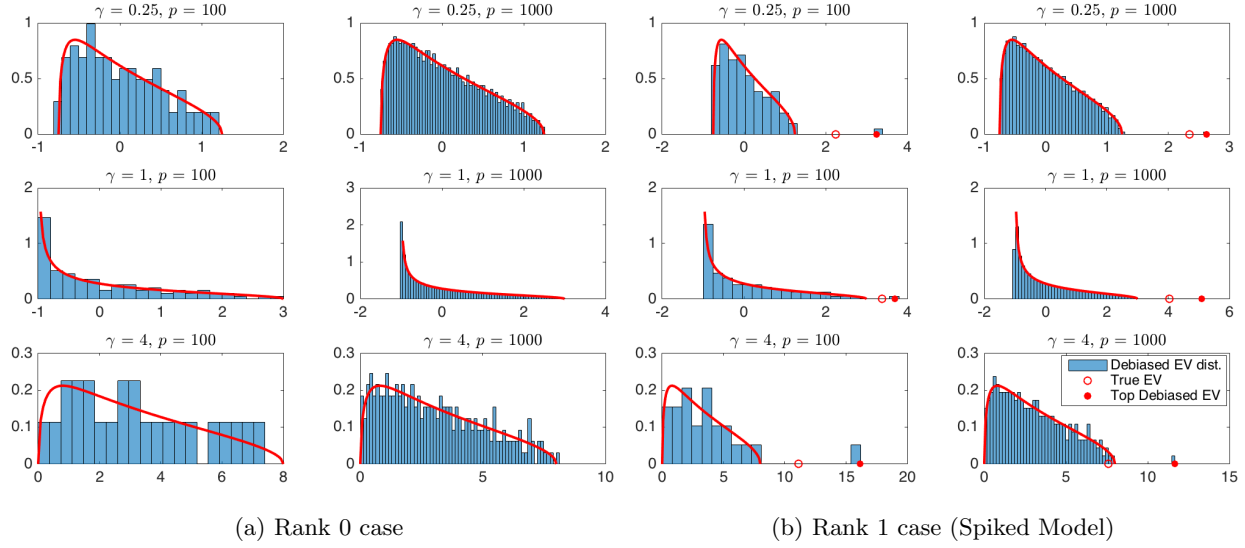


Figure 4.1: Empirical distribution of eigenvalues of whitened sample covariance S_w for different values of $\gamma = p/n$, with the corresponding Marchenko-Pastur density overlaid as a red curve. Data simulated according to 4.2. In the legend for (b), ‘Top Debiased EV’ refers top eigenvalue of S_w , while ‘True EV’ refers to the top eigenvalue of $D_n^{-1/2}\Sigma_x D_n^{-1/2}$, which we want to estimate.

that the standard Marchenko-Pastur law characterizes its limit spectrum. The standard Marchenko-Pastur distribution has a simple closed-form density, and there are many useful tools already available for low-rank covariance estimation (e.g., [Shabalin and Nobel, 2013](#); [Donoho et al., 2013](#)).

Theorem 4.2 (Marchenko-Pastur law for whitened covariance matrix). The eigenvalue distribution of $S_w + I_p$ converges almost surely to the standard Marchenko-Pastur distribution with aspect ratio γ .

In the proof presented in Appendix A.1.3, we deduce this from the Marchenko-Pastur law for the error matrix $n^{-1/2}\mathcal{E}$, for which standard results from [Bai and Silverstein \(2009\)](#) apply. The emergence of the standard MP law motivates the shrinkage method presented next.

4.2 Eigenvalue shrinkage

Since the early work of Stein ([Stein, 1956](#)) it is known that the estimation error of the sample covariance can be improved by eigenvalue shrinkage. Therefore, we will apply an eigenvalue shrinkage method to the whitened covariance matrix S_w . Let $\eta(\cdot)$ be a generic matrix shrinker, defined for symmetric matrices M with eigendecomposition $M = U\Lambda U^\top$ as $\eta(M) = U\eta(\Lambda)U^\top$. Here $\eta(\Lambda)$ is defined by applying the scalar shrinker η —typically a nonlinear scalar function—elementwise on the diagonal of the diagonal matrix Λ . Then our *whitened and shrunk* estimators will have the form

$$S_{w,\eta} = \eta(S_w) = \eta(D_n^{-1/2}S_d D_n^{-1/2}). \quad (6)$$

We are interested in settings where the clean signals lie on a low-dimensional subspace. We then expect the true covariance matrix Σ_x of the clean signals to be of low rank. However, based on Thm. 4.2, even in the case when $\Sigma_x = 0$, the empirical whitened covariance matrix is of full rank, and its eigenvalues have an asymptotic MP distribution. We are interested in shrinkers η that set all noise eigenvalues to zero, specifically $\eta(x) = 0$ for x within the support of the shifted MP distribution $x \in [(1 - \sqrt{\gamma})^2, (1 + \sqrt{\gamma})^2] - 1$. An example is operator norm shrinkage ([Donoho et al., 2013](#)).

Table 2: Spiked models: Summary of the original and whitened spiked model.

Model	Original	Whitened
Latent Signal	$X_i = u + z_i v$	$D^{-1/2} X_i = D^{-1/2} u + z_i D^{-1/2} v$
Marginal Covariance	$\text{Cov}[Y] = vv^\top + D$	$\text{Cov}[Y_w] = D^{-1/2} vv^\top D^{-1/2} + I_p$
Eigenvector	$v_{\text{norm}} = v/\ v\ $	$w = D^{-1/2} v/\ D^{-1/2} v\ $
Spike	$t = v^\top v$	$\ell = v^\top D^{-1} v$
SNR	$\frac{v^\top v}{\text{tr } D}$	$\frac{v^\top D^{-1} v}{p}$

However, whitening by $D_n \neq I_p$ also changes the direction of the eigenvectors. Therefore, to improve the accuracy of subspace estimates after eigenvalue shrinkage, we recolor multiplying back by the estimated standard errors. We define the *recolor* covariance estimator as:

$$S_r = D_n^{1/2} \cdot S_{w,\eta} \cdot D_n^{1/2}. \quad (7)$$

To understand whitening empirically, we perform two simulations. To generate non-negative i.i.d X_1, \dots, X_n lying in a low-dimensional space of dimension r , pick r vectors $v_1, v_2, \dots, v_r \in \mathbb{R}^p$ whose coordinates are i.i.d uniformly distributed in $[0, 1]$. For each i , sample r coefficients a_{i1}, \dots, a_{ir} independently from the uniform distribution on $[0, 1]$. Define $X_i = a_{i1}v_1 + \dots + a_{ir}v_r$. Note that X_i are non-negative, reside in a hyperplane spanned by v_1, \dots, v_r , and we can easily calculate the mean and covariance of X_i in terms of v_1, \dots, v_r . We normalize each direction v_1, \dots, v_r to have an L1 norm of unity. The coefficients a_{i1}, \dots, a_{ir} are also normalized so that $a_{i1} + \dots + a_{ir} = A$, where $A = 25(1 + \sqrt{\gamma})^2$ is a constant relating to signal strength, chosen empirically to push the spike outside of the bulk. Finally we sample $Y_i \sim \text{Poisson}_p(X_i)$ independently.

We display the histogram of the eigenvalues of the whitened covariance matrix S_w in one Monte Carlo instance, for $r = 0, 1$ and for several settings of γ and p on Figure 4.1. When $r = 0$, the standard MP distribution—shifted by -1 —is a good match, as can be seen on Fig. 4.1a. This is in accordance with Thm. 4.2. When $r = 1$, from Figure 4.1b, the standard MP distribution still matches the bulk of the noise eigenvalues. Moreover, we observe the same qualitative behaviour as in the classical spiked model, where the top *empirical eigenvalue* overshoots the *population eigenvalue*. In the next section we study this phenomenon more precisely.

4.2.1 The spiked model: Colored and whitened

To develop a method for estimating the eigenvalue after whitening and recoloring, we study a generalization of the spiked model (Johnstone, 2001) appropriate for our setting. Specifically, based on the covariance structure of the noisy signal, Eq. (2), we model the mean parameter $X = A'(\theta)$ of the exponential family—the clean observation—as a low rank vector. For simplicity, we will present the results in the rank one case, but they generalize directly to higher rank.

Suppose then that the i -th clean observation has the form $X_i = A'(\theta_i) = u + z_i v$, where u, v are deterministic p -dimensional vectors, and z_i are i.i.d. standardized random variables. In the Poisson case where $Y_i \sim \text{Poisson}_p(X_i)$, this assumes that the latent mean vectors are $X_i = u + z_i v$. The vector u is the global mean of the clean images, while v denotes the direction in which they vary.

For X_i to be a valid mean parameter for the exponential family, we need the additional condition that $u(j) + z_i|v(j)| \in A'(\Theta)$, for all i, j , where Θ is the natural parameter space of the exponential family, and $f(S)$ denotes the forward map of the set S under the function f . For instance, in the Poisson case, we need that $X_i(j) \geq 0$ for all i, j . If we take z_i to be uniform random variables on $[-\sqrt{3}, \sqrt{3}]$, so that their variance is unity, then a sufficient condition is that $u(j) \geq \sqrt{3}|v(j)|$ for all j .

Using our formula for the marginal covariance of the noisy observations, $\text{Cov}[Y] = \text{Cov}[X] + \mathbb{E} \text{diag}[V(X)]$,

and defining $D = \mathbb{E} \text{diag}[V(X)]$, we obtain

$$\text{Cov}[Y] = vv^\top + D. \quad (8)$$

For instance, in the Poisson case we have $\text{Cov}[Y] = vv^\top + \text{diag}[u]$.

We whiten the observations dividing by the elements of $D^{1/2}$. The elements of D are expected values of variances. They are thus positive, except for coordinates that can be discarded because they have no variability. The whitened observations are $Y_w = D^{-1/2}Y$, and their population covariance matrix is

$$\text{Cov}[Y_w] = D^{-1/2}vv^\top D^{-1/2} + I_p. \quad (9)$$

We now compare this with the usual *standard spiked model* (Johnstone, 2001) where the observations Y_w are Gaussian and have covariance matrix

$$\text{Cov}[Y_w] = \ell ww^\top + I_p,$$

where $\ell \geq 0$ and the vector w has unit norm. This model has been thoroughly studied in probability theory and statistics. In particular, the Baik-Ben Arous-Péché (BBP) phase transition (PT) (Baik et al., 2005) shows that when $n, p \rightarrow \infty$ such that $p/n \rightarrow \gamma > 0$, the top eigenvalue of the sample covariance matrix asymptotically separates from the Marchenko-Pastur bulk if the population spike $\ell > \sqrt{\gamma}$. Otherwise, the top sample eigenvalue does not separate from the MP bulk. This was shown first for complex Gaussian observations, then generalized to other distributions (see e.g., Paul and Aue, 2014; Yao et al., 2015).

Heuristically, comparing with (9), we surmise that a spiked model with $\ell = v^\top D^{-1}v$ and $w = D^{-1/2}v / \|D^{-1/2}v\|$ is a good approximation in our case. In particular the BBP phase transition should happen approximately when

$$v^\top D^{-1}v = \sqrt{\gamma}.$$

For instance in the Poisson case, the condition is $v^\top \text{diag}[u]^{-1}v = \sqrt{\gamma}$. In the next sections, we provide numerical evidence for this surmise, and develop its consequences.

4.2.2 Whitening improves SNR

In this section we justify our whitening method theoretically, showing that it can improve the signal-to-noise ratio. This was observed empirically in previous work on covariance estimation in a related setting, but a theoretical explanation was lacking (Bhamre et al., 2016).

As usual, we define the the signal-to-noise ratio (SNR) of a “signal+noise” vector observation $y = s + n$ as the ratio of the trace of the covariances of s and of n . In the unwhitened model from Eq. (8), the SNR equals

$$\frac{\text{tr Cov}[X]}{\text{tr } \mathbb{E} \text{diag}[V(X)]} = \frac{\text{tr } vv^\top}{\text{tr } D} = \frac{v^\top v}{\text{tr } D}.$$

In particular, the SNR is of order $O(1/p)$ in the typical case when the vector v has norm of unit order. In the whitened model from Eq. (9), the SNR equals $v^\top D^{-1}v/p$.

Suppose now that v is approximately delocalized in the sense that

$$p \cdot v^\top D^{-1}v \approx \text{tr } D^{-1} \cdot v^\top v.$$

This holds for instance if the entries of v are i.i.d. centered random variables with the same variance σ^2 . In that case, $\mathbb{E}v^\top D^{-1}v = \sigma^2 \text{tr } D^{-1}$ and $\mathbb{E}v^\top v = \sigma^2 p$, and under higher moment assumptions it is easy to show the concentration of these quantities around their means, showing delocalization as above. If v is delocalized, then we obtain that the SNR in the whitened model is higher than in the original model. Indeed, this follows because D is diagonal, so by the Cauchy-Schwarz inequality

$$\frac{v^\top D^{-1}v}{p} \approx \frac{\text{tr } D^{-1} \cdot v^\top v}{p^2} = \frac{\sum_{i=1}^p D_i^{-1} \cdot v^\top v}{p^2} \geq \frac{v^\top v}{\sum_{i=1}^p D_i} = \frac{v^\top v}{\text{tr } D}.$$

Moreover, we can define the improvement (or amplification) in SNR as

$$\mathcal{I} = \frac{\text{tr } D}{p} \cdot \frac{v^\top D^{-1} v}{v^\top v}. \quad (10)$$

The above heuristic can be formalized into a rigorous result as follows:

Proposition 4.3. Suppose the signal eigenvector v is delocalized in the sense that for some $\varepsilon > 0$,

$$\frac{v^\top D^{-1} v}{v^\top v} \geq (1 - \varepsilon) \frac{\text{tr}[D^{-1}]}{p}.$$

Let moreover

$$\beta = \frac{\sum_{i=1}^p D_i \cdot \sum_{i=1}^p D_i^{-1}}{p^2} \geq 1.$$

Then the SNR is improved by whitening, by a ratio $\mathcal{I} \geq (1 - \varepsilon)\beta$.

If β is large and $\varepsilon > 0$ is small, the SNR can improve substantially.

4.2.3 Eigenvalue shrinkage and scaling

We now continue with our overall goal of estimating the covariance matrix $\text{Cov}[X] = vv^\top$ of X in the spiked model. The covariance matrix has one nonzero eigenvalue $t = \|v\|^2$ and corresponding eigenvector $v_{\text{norm}} = v/\|v\|$. Starting with the recolored covariance matrix S_r , we use its top eigenvector as an estimator of v_{norm} . To estimate t , a first thought is to use the top empirical eigenvalue of S_r . However, it turns out that this naive estimator is biased. To understand and correct the bias, we review some basic facts about PCA in high dimensions.

In the case of independent data with identical variances, the cumulative work of many authors (e.g., [Baik et al., 2005](#); [Baik and Silverstein, 2006](#); [Paul, 2007](#); [Benaych-Georges and Nadakuditi, 2011](#), etc) shows that if the population spike is above the BBP phase transition—i.e., $\ell > \sqrt{\gamma}$ —then the top sample spike pops out from the Marchenko-Pastur bulk describing the distribution of the “noise” eigenvalues. The top eigenvalue will converge to the value given by *the spike forward map*:

$$\lambda(\ell; \gamma) = \begin{cases} (1 + \ell) \left(1 + \frac{\gamma}{\ell}\right) & \text{if } \ell > \gamma^{1/2}, \\ (1 + \gamma^{1/2})^2 & \text{otherwise.} \end{cases}$$

We conjecture that the BBP phase transition also applies to our case, and describes the behavior of the spikes after whitening. We have verified this in numerical simulations in certain cases (data not shown). Therefore, as in many previous works, we propose to estimate ℓ consistently by inverting the spike forward map (see e.g., [Lee et al., 2010](#); [Donoho et al., 2013](#)), i.e., defining $\hat{\ell} = \lambda^{-1}(\lambda_{\max}(S_w))$. [Donoho et al. \(2013\)](#) provided an asymptotic optimality result for this estimator of the spike in operator norm loss.

Once we have a good estimator $\hat{\ell}$ of $\ell = v^\top D^{-1} v$, a first thought is to estimate $t = v^\top v$ as the top eigenvalue of the recolored covariance matrix S_r . However, we will show that this estimator is biased, and we will propose a suitable bias-correction.

The estimation accuracy is affected in a significant way by the inconsistency of the empirical eigenvector \hat{w} of S_w as an estimator of the true eigenvector $w = D^{-1/2} v / \|D^{-1/2} v\|$. We can quantify this heuristically based on results for Gaussian data. In the Gaussian standard spiked model the empirical and true eigenvectors have an asymptotically deterministic angle: $(w^\top \hat{w})^2 \rightarrow c^2(\ell; \gamma)$ almost surely, where $c(\ell; \gamma)$ is the cosine forward map given by (e.g., [Paul, 2007](#); [Benaych-Georges and Nadakuditi, 2011](#), etc):

$$c(\ell; \gamma)^2 = \begin{cases} \frac{1 - \gamma/\ell^2}{1 + \gamma/\ell} & \text{if } \ell > \gamma^{1/2}, \\ 0 & \text{otherwise.} \end{cases}$$

Heuristically, in finite samples we can write

$$\hat{w} \approx cw + s\varepsilon,$$

where $s = s(\ell; \gamma) \geq 0$ is the sine defined by $s^2 = 1 - c^2$, and ε is white noise with approximate norm $\|\varepsilon\| = 1$. Then, since $w^\top Dw = v^\top v / v^\top D^{-1}v = t/\ell$, and $\varepsilon^\top D\varepsilon \approx \text{tr}(D)/d$, we have

$$\begin{aligned} \|\hat{v}\|^2 &\approx \ell \cdot \hat{w}^\top D \hat{w} \approx \ell \cdot (cw + s\varepsilon)^\top D (cw + s\varepsilon) \approx \ell \cdot (c^2 w^\top D w + s^2 \varepsilon^\top D \varepsilon) \\ &\approx tc^2 + \ell s^2 \text{tr}(D)/p. \end{aligned}$$

Comparing this to $\|v\|^2 = t = tc^2 + ts^2$, we find that the bias is

$$\|\hat{v}\|^2 - t \approx s^2 (v^\top D^{-1}v \cdot \text{tr}(D)/p - v^\top v) = s^2 t \cdot (\mathcal{I} - 1) \geq 0.$$

This suggests that $\|\hat{v}\|^2$ is an upward biased estimator of $t = \|v\|^2$. Interestingly, the bias is closely related to the improvement \mathcal{I} in SNR.

To correct the bias, we propose an estimator of the form $\hat{t}(\alpha) = \alpha \|\hat{v}\|^2$ for which $\alpha \|\hat{v}\|^2 \approx \|v\|^2$. We have $\|\hat{v}\|^2 \approx t \cdot [1 + s^2(\mathcal{I} - 1)]$, suggesting that we define $\alpha = [1 + s^2(\mathcal{I} - 1)]^{-1}$. This quantity is an unknown population parameter, and it depends on s^2 and \mathcal{I} . We can estimate s^2 in the usual way by $\hat{s}^2 = s^2(\hat{\ell}; \gamma)$. Since \mathcal{I} itself depends on the parameter t we are trying to estimate, we plug in the same estimator $\hat{t}(\alpha) = \alpha \|\hat{v}\|^2$, leading to the following estimator of \mathcal{I} (where we also define τ for future use):

$$\hat{\mathcal{I}}(\alpha) = \frac{\text{tr } D_n}{p} \cdot \frac{\hat{\ell}}{\hat{t}(\alpha)} = \frac{\text{tr } D_n}{p} \cdot \frac{\hat{\ell}}{\alpha \|\hat{v}\|^2} = \frac{\tau}{\alpha}.$$

Since $\alpha = [1 + s^2(\mathcal{I} - 1)]^{-1}$, it is reasonable to require that the following fixed-point equation holds for α :

$$\hat{\alpha} = [1 + \hat{s}^2(\hat{\mathcal{I}}(\hat{\alpha}) - 1)]^{-1}.$$

We can equivalently rewrite the fixed-point equation as

$$\frac{1}{\hat{\alpha}} = \hat{c}^2 + \hat{s}^2 \hat{\mathcal{I}}(\hat{\alpha}) = \hat{c}^2 + \hat{s}^2 \frac{\tau}{\hat{\alpha}}.$$

Equivalently, when $\hat{c}^2 > 0$,

$$\hat{\alpha} = \frac{1 - \hat{s}^2 \tau}{\hat{c}^2}. \quad (11)$$

When $\hat{c}^2 = 0$, i.e., when $\hat{\ell} \leq \sqrt{\gamma}$, the equation reads $1/\hat{\alpha} = \tau/\hat{\alpha}$. If $\tau = 1$, this has solution $\alpha = 1$, else it has no solution. Therefore, when $\hat{c}^2 = 0$, we define $\hat{\alpha} = 1$. We finally define $\hat{t}(\hat{\alpha}) = \hat{\alpha} \|\hat{v}\|^2$. The implication is that we ought to rescale the estimated magnitude of the signal subspace corresponding to v by $\hat{\alpha}$.

In the multispiked case, suppose $X_j = u + \sum_{i=1}^r z_{ij} v_i$. Then the marginal covariance of Y is $\text{Cov}[Y] = \sum_{i=1}^r v_i v_i^\top + D$. Suppose that the v_i are sorted in the order of decreasing norm. Suppose moreover that the recolored sample covariance S_r has the form

$$S_r = \sum_{i=1}^r \hat{v}_i \hat{v}_i^\top = \sum_{i=1}^r \hat{\lambda}_i \hat{u}_i \hat{u}_i^\top,$$

where \hat{u}_i are orthonormal, and the $\hat{\lambda}_i \geq 0$ are sorted in decreasing order. Based on our above discussion, we define the *scaled* covariance matrix as

$$S_s = \sum_{i=1}^r \hat{\alpha}_i \hat{v}_i \hat{v}_i^\top, \quad (12)$$

where $\hat{\alpha}_i$ is defined in (11), with $\hat{s}^2 = \hat{s}_i^2 = s^2(\hat{\ell}_i; \gamma)$. This concludes our methodology for covariance estimation. We use the terminology *ePCA* for the eigendecomposition of the covariance matrix estimator

Algorithm 1: Covariance matrix estimation and *e*PCA

Input: data $Y = [Y_1, \dots, Y_n]^\top \in \mathbb{R}^{n \times p}$, desired rank $r \leq p$,
mean-variance map V of exponential family

Output: covariance estimator $S_s \in \mathbb{R}^{p \times p}$ of noiseless vectors; *e*PCA: eigendecomposition of S_s

- 1 Sample mean $\bar{Y} \leftarrow n^{-1} \sum_{i=1}^n Y_i$
 - 2 Sample covariance matrix $S \leftarrow n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^\top$
 - 3 Variance estimates $D_n \leftarrow \text{diag}[V(\bar{Y})]$
 - 4 Whitening $S_w \leftarrow D_n^{-1/2} S D_n^{-1/2}$
 - 5 Eigendecomposition $S_w = \hat{W} \Lambda \hat{W}^\top$
 - 6 Eigenvalue shrinkage $S_{w,\eta} = \hat{W} \eta(\Lambda_r) \hat{W}^\top = \sum_{i=1}^r \hat{\ell}_i \hat{w}_i \hat{w}_i^\top$ of top r eigenvalues $\Lambda_r = \text{diag}(\lambda_1, \dots, \lambda_r)$
 - 7 Scaling coefficients $\hat{\alpha}_i = [1 - s^2(\hat{\ell}_i; \gamma) \tau_i] / c^2(\hat{\ell}_i; \gamma)$ (as in (11))
 - 8 Recoloring $S_r = D_n^{1/2} S_{w,\eta} D_n^{1/2}$
 - 9 Scaling $S_s = \sum \hat{\alpha}_i \hat{v}_i \hat{v}_i^\top$, where eigendecomposition of $S_r = \sum \hat{v}_i \hat{v}_i^\top$
-

(12). Both the eigenvalues and the eigenvectors of this estimator are different from those of the sample covariance matrix.

*e*PCA is summarized in Alg. 1. As discussed at the beginning of Sec. 4.2, we assume here that we have a guess r for the number of PCs. In exploratory analyses, one can often try several choices for r . While there are many formal methods for choosing the rank r (see e.g., Jolliffe, 2002), it is beyond our scope to investigate them in detail here (see Sec. 7).

4.2.4 Simulations with *e*PCA

We report the results of a simulation study with *e*PCA. We simulate data Y_i from the Poisson model $Y_i \sim \text{Poisson}_p(X_i)$, where the mean parameters are $X_i = u + z_i \ell^{1/2} v$, the z_i are i.i.d. unit variance random variables uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$, and $u \in \mathbb{R}^p$ has entries $u(i)$ sorted in increasing order on a uniform grid on $[1, 3]$, while $v \in \mathbb{R}^p$ has entries $v(i)$ sorted in increasing order on a uniform grid on $[-1, 1]$, standardized so that $\|v\|^2 = 1$. We take the dimension $p = 500$, and $\gamma = 1/2$, so $n = 1000$. The phase transition occurs when the spike is $\ell = \sqrt{\gamma}/v^\top \text{diag}[u]^{-1} v \approx 1.2$. We vary the spike strength ℓ on a uniform grid of size 20 on $[0, 3]$. We generate $n_M = 100$ independent Monte Carlo trials, and compute the mean of the recolored spike estimator $\hat{t} = \|\hat{v}\|^2$ and the *e*PCA—or scaled—estimator $\hat{t}(\hat{\alpha}) = \hat{\alpha} \|\hat{v}\|^2$ over these trials.

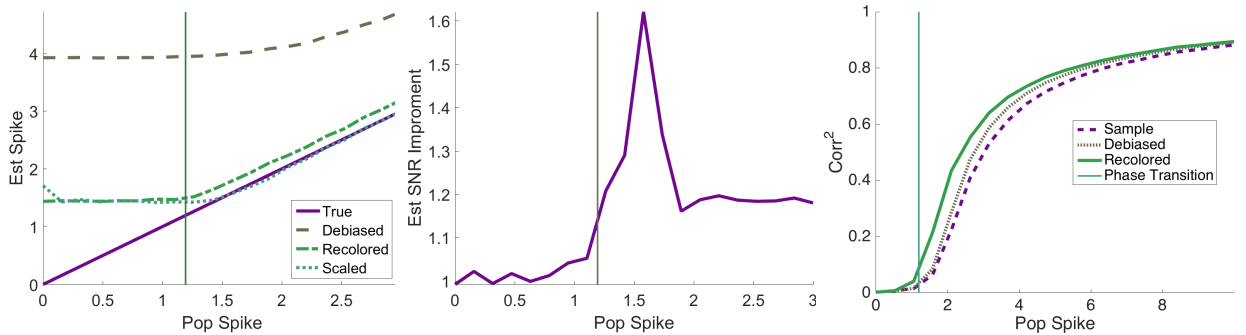


Figure 4.2: MC simulation with *e*PCA. Left: Spike estimation; true, debiased, recolored, and scaled (*e*PCA) spike estimators. Middle: Estimated improvement in SNR due to whitening. Right: Squared correlation between the true signal vector v and leading eigenvector of various covariance estimates; sample, debiased, recolored (*e*PCA). Plotted against the spike, which is directly proportional to SNR.

The results displayed in Fig. 4.2 (left) show that the e PCA/scaled estimator (top eigenvalue of S_s) reduces the bias of the recolored estimator (top eigenvalue of S_r) especially for large spikes. Both are much better than the debiased estimator (top eigenvalue of S_d). Below the phase transition (vertical line), both estimators have the same approximate value.

We can also define an estimator of the improvement in SNR \mathcal{I} , as $\hat{\mathcal{I}}(\hat{\alpha})$. The mean of this estimator over the same simulation is displayed in Fig. 4.2 (middle). We observe that it is approximately unity below the phase transition in the whitened model. This makes sense, because the spike is below the PT both before and after whitening. The improvement in SNR has a “jump” just above the PT, because the spike pops out from the bulk after whitening. This is where whitening helps the most. However, $\hat{\mathcal{I}}$ is not “infinitely large”, because the signal is detectable in the unwhitened spectrum, except it is spread across all eigenvalues (see e.g., Dobriban, 2016). Finally, $\hat{\mathcal{I}}(\hat{\alpha})$ drops to a lower value, still above unity, and stabilizes. We find this an illuminating way to quantify the improvement due to whitening.

Finally, we also display the mean of the squared correlation between the true and empirical eigenvectors of various covariance matrix estimators in figure 4.2 (right). The predicted PT matches the empirical PT. The e PCA eigenvector—top eigenvector of S_s —in this case agrees with the eigenvector of the recolored covariance matrix S_r , because both are of rank one. e PCA has the highest correlation, and the improvement is significant just above the PT.

4.3 Whitening agrees with HWE normalization

It is of special interest that for Binomial(2) data, and specifically for biallelic genetic markers such as Single Nucleotide Polymorphisms, our whitening method recovers exactly the well-known normalization assuming Hardy-Weinberg equilibrium (HWE). In these datasets the entries X_{ij} are counts ranging from 0 to 2 denoting the number of copies of the variant allele of biallelic marker j in the genome of individual i . The HWE normalization divides the entries of SNP j by $\sqrt{2\hat{p}_j(1-\hat{p}_j)}$, where $\hat{p}_j = (2n)^{-1} \sum_i X_{ij}$ is the estimated allele frequency of variant j (e.g., Patterson et al., 2006, p. 2075). This is exactly the same as our whitening method assuming that the individual data points X_{ij} are Binomial(2)-distributed (see Sec. A.2).

Previously, the HWE normalization was motivated by a connection to genetic drift, and by the empirical observation that it improves results on observational and simulated data (Patterson et al., 2006, p. 2075). Our theoretical results provide justification for using HWE normalization. In particular, our result on the Marchenko-Pastur law (Thm. 4.2) suggests that the Marchenko-Pastur is an accurate null distributions after whitening. Numerical results reported in Sec. A.3 also suggest that the approximations to both the MP law and the Tracy-Widom distribution for the top eigenvalue are more accurate than after standardization. In addition, our result on the improved SNR (Prop. 4.3) suggests that “signal” components become easier to identify after whitening.

However, in practice we often see similar results using whitening and standardization. In many SNP datasets, the variants not approximately in HWE—i.e., the variants for which a goodness of fit test to a Binomial(2) distribution is rejected—are removed as part of data quality control. Therefore, most remaining SNPs have an empirical distribution well fit by a Binomial(2). In such cases standardization and whitening lead to similar results.

4.4 Multiple exponential families

Our methodology can handle observations for which the different coordinates have different exponential family noise. For instance, some coordinates may be Gaussian, others may be Binomial, and yet others may be Poisson. This is important in applications where heterogeneous datasets are integrated. For instance, in genomics, there are many different data types, including Gaussian (log-gene expression level), Binomial count (SNP), and Poisson/negative Binomial (RNA-seq). Our methodology extends because it only depends on the first two moments of the distributions, while our theoretical results only depend on the analytic properties of the coordinate-wise mean-variance map V . Thus both the methods and the theory extend to heterogeneous data.

5 Denoising

As an application of *ePCA*, we develop a method to denoise the observed data. Formally the goal of denoising is to predict the noiseless signal vectors $X_i = A'(\theta_i)$. Our model is a random effects model (see e.g., [Searle et al., 2009](#)), hence we will predict X_i using the Best Linear Predictor—or BLP ([Searle et al., 2009](#), Sec. 7.4). Let $\mathbb{E}(X|Y) = BY + C$ denote the linear predictor of the random vector X using Y with minimum MSE, where B is a deterministic matrix, and C is a deterministic vector. This is known under various names in statistics and signal processing, including the “Wiener filter”, see Sec. 1.4. We will refer to it as the BLP, which is the common terminology in random effects models. It is well known (e.g., [Searle et al., 2009](#), Sec. 7.4) that

$$B = \Sigma_x [\text{diag}[\mathbb{E}A''(\theta)] + \Sigma_x]^{-1} \text{ and } C = \text{diag}[\mathbb{E}A''(\theta)] [\text{diag}[\mathbb{E}A''(\theta)] + \Sigma_x]^{-1} \mathbb{E}A'(\theta).$$

The BLP depends on the unknown parameters Σ_x , $\text{diag}[\mathbb{E}A''(\theta)]$, and $\mathbb{E}[A'(\theta)]$. The standard strategy, known as *Empirical BLP* or EBLP (e.g., [Searle et al., 2009](#)) is to estimate these unknown parameters using the entire dataset, and denoise the vectors Y_i by plug-in:

$$\hat{X}_i = \hat{\Sigma}_x \left[\text{diag}[\hat{\mathbb{E}}A''(\theta)] + \hat{\Sigma}_x \right]^{-1} Y_i + \text{diag}[\hat{\mathbb{E}}A''(\theta)] \left[\text{diag}[\hat{\mathbb{E}}A''(\theta)] + \hat{\Sigma}_x \right]^{-1} \bar{Y}.$$

We will use *ePCA*, i.e., the scaled recolored covariance matrix S_s proposed in (12) to estimate Σ_x . As before in Sec. 3.2, we will use the sample mean \bar{Y} to estimate $\mathbb{E}[A'(\theta)]$, and $V(\bar{Y})$ to estimate the noise variances $\mathbb{E}A''(\theta)$. However, in principle different estimators could be used.

Concrete examples include:

- For normal distributions, the posterior mean formula simplifies to $\tilde{\mathbb{E}}(X|Y) = (\sigma^2 I_p + \Sigma_x)^{-1} \Sigma_x Y + (\sigma^2 I_p + \Sigma_x)^{-1} \sigma^2 m$. This is the most well-known form of the Wiener filter.
- For the Poisson distribution, we get $\tilde{\mathbb{E}}(X|Y) = \Sigma_x [\text{diag}[\mathbb{E}X] + \Sigma_x]^{-1} Y + \text{diag}[\mathbb{E}X] [\text{diag}[\mathbb{E}X] + \Sigma_x]^{-1} \mathbb{E}X$. Thus, the EBLP denoiser becomes

$$\hat{X}_i = S_s (\text{diag}[\bar{Y}] + S_s)^{-1} \hat{Y}_i + \text{diag}[\bar{Y}] (\text{diag}[\bar{Y}] + S_s)^{-1} \bar{Y}.$$

In some examples there are coordinates where $\bar{Y}(j) = 0$. For instance in our XFEL application this corresponds to pixels where no photon was observed during the entire experiment. This causes a problem because the matrix $\hat{\Sigma} = \text{diag}[\bar{Y}] + S_s$ may no longer be invertible: S_s is of low rank, while $\text{diag}[\bar{Y}]$ is also not of full rank. To avoid this problem, we implement a ridge-regularized covariance estimator $\hat{\Sigma}_\varepsilon = (1 - \varepsilon)\hat{\Sigma} + \varepsilon \cdot \tilde{m}I_p$ as in [Ledoit and Wolf \(2004\)](#), where $\tilde{m} = \text{tr} \hat{\Sigma}/p$ and $\varepsilon > 0$ is a small constant. Note that $\text{tr} \hat{\Sigma}_\varepsilon = \text{tr} \hat{\Sigma}$. The ridge-regularized estimator $\hat{\Sigma}_\varepsilon$ has a small bias, but is invertible. In our default implementation we take $\varepsilon = 0.1$. The same method can be implemented for any exponential family.

6 Experiments

We apply *ePCA* to a simulated XFEL dataset, and an empirical genetics dataset, comparing with standard PCA.

6.1 XFEL images

We simulate 20,000 noiseless XFEL diffraction intensity maps of a lysozyme (Protein Data Bank 1AKI) with Condor ([Hantke et al., 2016](#)). We rescale the average pixel intensity to 0.04 such that shot noise dominates, following previous work (e.g., [Schwander et al., 2012](#)). To sample an arbitrary number of noisy diffraction patterns, we sample an intensity map at random, and then sample the photon count of each detector pixel from a Poisson distribution whose mean is the pixel intensity. The images are 64 pixels by 64 pixels, so $p = 4096$. Figure 1.1 illustrates the intensity maps and the resulting noisy diffraction patterns.

6.1.1 Covariance estimation

We report the error of covariance estimation as it varies with sample size in figure 6.1. Each of the diagonally debiased, recolored (a.k.a., debiased shrunk), and scaled (a.k.a., debiased shrunk and eigenvalue bias-corrected) covariance estimates S_d , S_r , S_s improves on the sample covariance S . The largest improvement is due to the diagonal debiasing. For spectral norm, it is unclear if the recoloring and scaling steps decrease the MSE in this case. For Frobenius norm, however, they clearly improve the MSE, though this improvement diminishes with larger sample size. In addition, the low variance of the estimation error across Monte Carlo trials suggests the stability of our method.

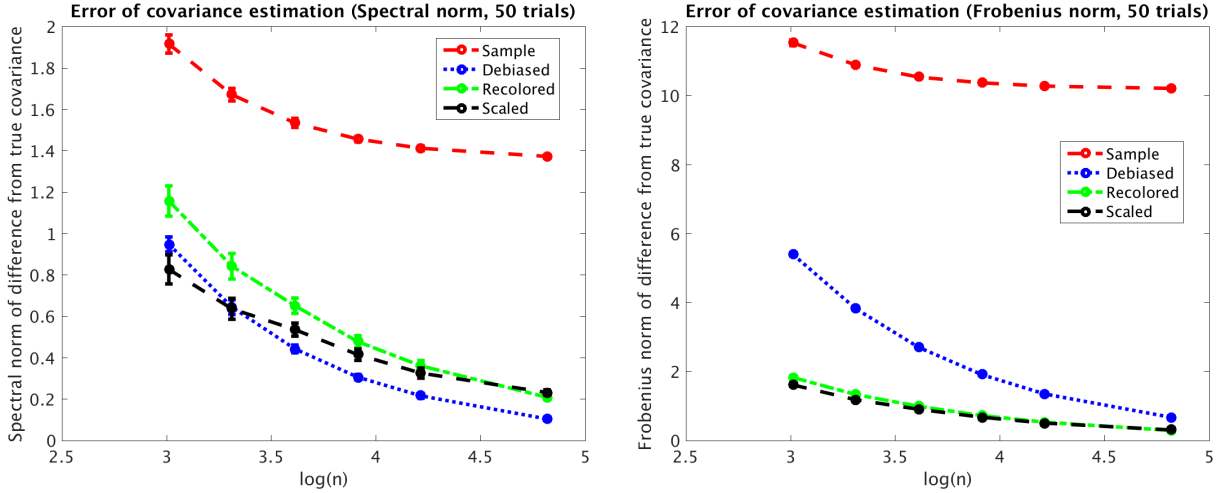


Figure 6.1: Error of Covariance matrix estimation, measured as the spectral norm (left) and Frobenius norm (right) of the difference between each covariance estimate (Sample, Debiased, Recolored, Scaled) and the true covariance matrix. For each covariance estimate, we fixed the rank to be $r = 12$. Other values of r led to similar results.

Figure 6.2 summarizes the error of eigenvalue estimation. The e PCA eigenvalues are indeed much closer to the true eigenvalues than the eigenvalues of the debiased or sample covariance matrices S_d or S . Impressively, the estimation error for e PCA eigenvalues is small regardless of sample size.

We visualize the eigenvectors (or eigenimages) for XFEL diffraction patterns in Figure 6.3. As can be seen, the recolored method accurately estimates two more eigenimages with small eigenvalues than alternative methods. This shows that recoloring significantly improves the denoising of XFEL data.

We note that e PCA/recolored eigenvectors 1 to 3 in Figure 6.3 appear misaligned with the corresponding true eigenvectors. A likely explanation is that the top eigenvectors have similar eigenvalues, leading to some reordering and rotation in the estimated eigenvectors. This affects the alignment of the estimated eigenvectors and the true eigenvectors. Therefore, we also report the error of estimating the overall low-rank subspace, for fixed rank $r = 12$, measured as the estimation MSE of the projection matrix $U_r U_r^T$. Other values of r lead to comparable performance. Figure 6.4a clearly shows that the e PCA/recolored covariance matrix best estimates the low-rank subspace inhabited by the clean data. Moreover, it is significantly more efficient statistically—the e PCA/recolored subspace estimate improves faster with more data than the two alternative estimates.

As a special note, our final scaling step S_s does not change the eigenvectors compared to recoloring S_r , so the recoloring and scaling are equivalent for all purely eigenvector-dependent measures of error (such as subspace estimation). In those cases recoloring is identical to e PCA.

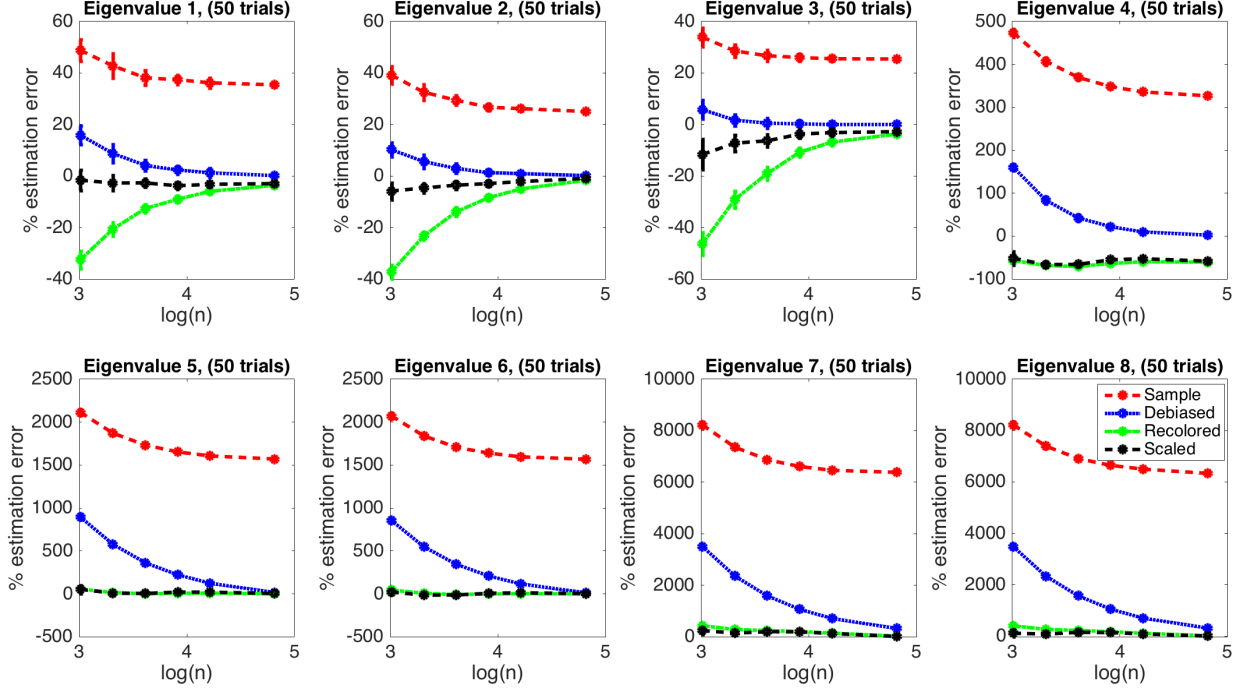


Figure 6.2: Error of eigenvalue estimation for the top 8 eigenvalues, measured as percentage error relative to the true eigenvalue, for XFEL data. We plot the mean and standard deviation (as error bars) over 50 Monte Carlo trials.

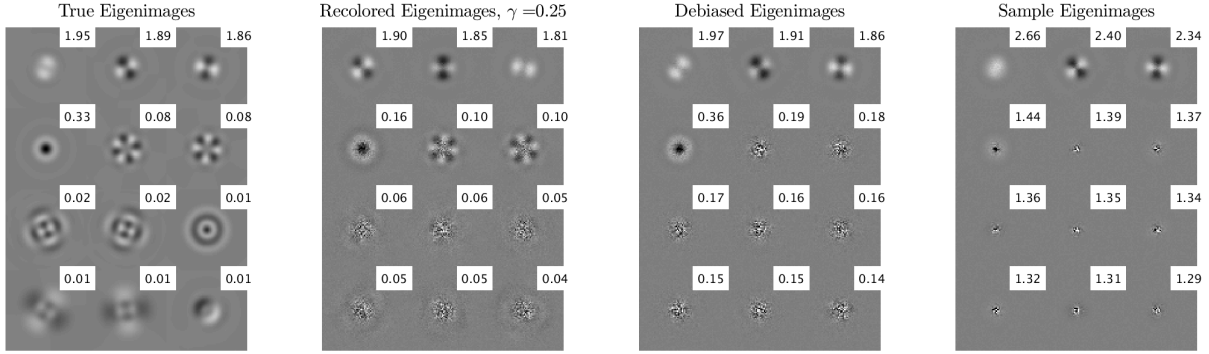


Figure 6.3: XFEL Eigenimages for $\gamma = 1/4$

6.1.2 Denoising

Finally, we report the results of denoising the XFEL patterns. We compare “PCA denoising”, i.e., orthogonal projection onto sample or e PCA/recolored “eigenimages”, and EBLP denoising. PCA denoising results in clear artifacts that are exacerbated in the high dimensional regime, while the reconstructions after EBLP denoising are always the closest to the clean images (Fig. 6.5). In EBLP denoising, our scaled covariance matrix leads to much better results than the sample covariance matrix. This underscores that it is important to bias-correct before plugging in to the optimal BLP formulas. The EBLP denoiser is also better than both alternatives as measured by mean squared error $pn^{-1} \sum_{i=1}^n \|\hat{Y}_i - X_i\|^2$, see Fig. 6.4b. Notably, the advantage

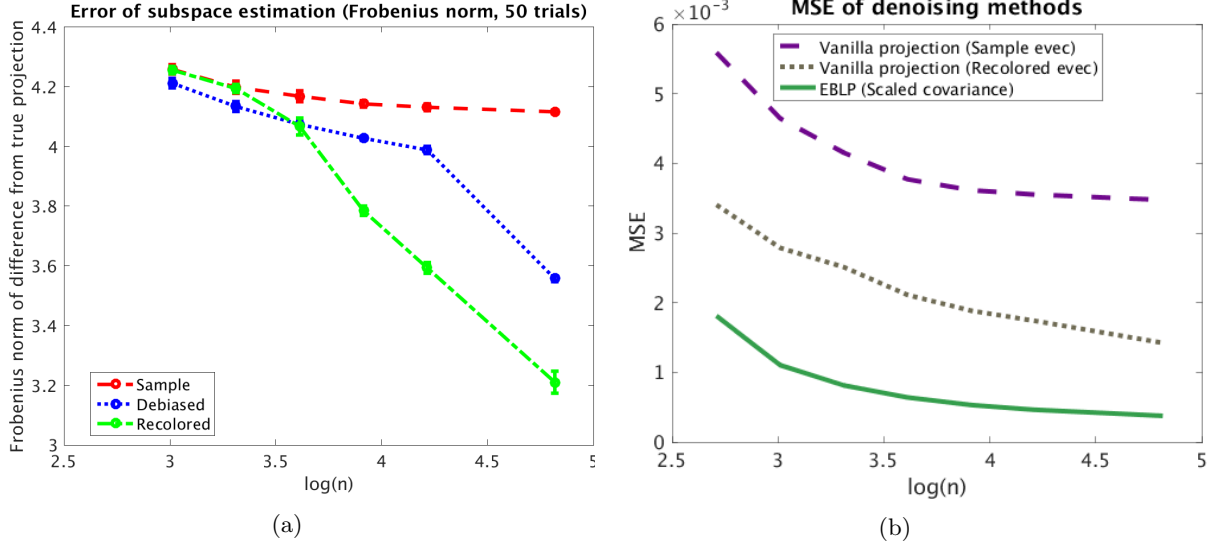


Figure 6.4: a) Subspace estimation error for XFEL data. We plot the mean and standard deviation (as error bars) over 50 Monte Carlo trials. b) Evaluation of denoising methods by the reconstruction error of clean images, against sample size (\log_{10} scale)

of EBLP is greater in the high dimensional regime.

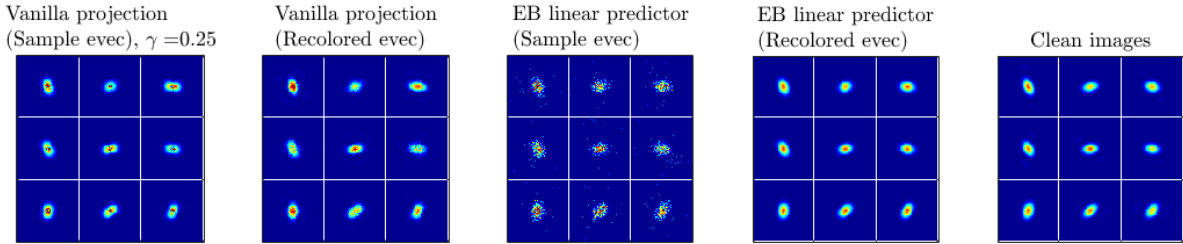


Figure 6.5: Sampled reconstructions using the XFEL dataset ($n = 16,384$; $p = 4096$), fixing the rank of covariance estimates at $r = 12$

6.2 HGDP dataset

We also apply e PCA to a subset of the Human Genome Diversity Project (HGDP) dataset (Li et al., 2008). The HGDP dataset contains Single Nucleotide Polymorphism (SNP) markers obtained from human samples across the globe. We obtained a homogeneous random set of $n = 20$ Caucasian samples from the CEU cohort, typed on $p = 120,631$ SNPs. We removed SNPs that showed no variability, with $p' = 107,026$ SNPs remaining. For each SNP we imputed missing data as the mean of the available samples. We then computed the PC scores starting from two covariance matrices: (1) the one obtained after usual standardization of each feature to have unit norm, and (2) S_w obtained by using our whitening method, which in this case agrees with HWE normalization as defined in e.g., Patterson et al. (2006) (see Sec. 4.3).

In Fig. 6.6 we see that whitening/HWE normalization appears to lead to a clearer structure in the PC scores than standardization. There are two samples on the PC scores after standardization that appear to be extreme outliers. However, our data is a homogeneous random sample, so we do not expect any outliers.

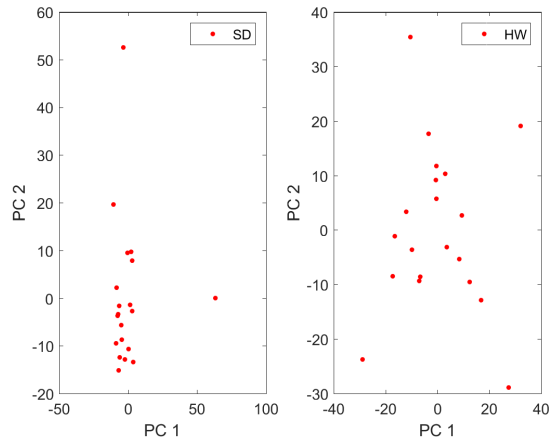


Figure 6.6: HGDP dataset: PC scores of 20 CEU samples after standardization (SD, left) and whitening/HWE normalization (HW, right).

This suggests that standardization is more sensitive to outliers or artifacts. However, the difference between the two methods is relatively small. These empirical results reinforce our theory, and are in line with the existing empirical observations about the superiority of HWE normalization (Patterson et al., 2006).

7 Future work

In the context of XFEL imaging, each diffraction pattern is equally likely to appear in any possible in-plane rotation. As a result, the covariance matrix commutes with rotations and is block diagonalized in any basis made of outer products of radial functions and angular Fourier modes, such as the Fourier-Bessel basis (Zhao et al., 2016). Indeed, the eigenimages in Figure 6.2 clearly show angular oscillation. Incorporating “steerability” into our methodology, that is, including the block diagonal structure into the estimation framework would lead to more accurate covariance estimation, as it effectively reduces the dimension p .

Furthermore, it would be valuable to prove rigorously the results about the spiked model for exponential families. Our results in Section 4.1.1 only cover the null case, but it would be useful to know rigorously the behavior of the signal eigenvalues in non-null cases. Finally, it could be useful to have a principled method to choose the rank.

8 Acknowledgements

The authors wish to thank Joey Arthur, Yuval Kluger, Filipe Maia, Nick Patterson, Kris Sankaran, Joel Tropp, Ramon van Handel, Teng Zhang, and Jane Zhao for valuable discussions and for help with software and data.

A. S. was partially supported by Award Number R01GM090200 from the NIGMS, FA9550-12-1-0317 from AFOSR, Simons Foundation Investigator Award and Simons Collaborations on Algorithms and Geometry, and the Moore Foundation Data-Driven Discovery Investigator Award. E. D. was partially supported by NSF grant DMS-1407813, and by an HHMI International Student Research Fellowship.

References

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley New York, 2003.

- Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, 2009.
- J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.
- J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697, 2005.
- F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.
- T. Bhamre, T. Zhang, and A. Singer. Denoising and covariance estimation of single particle cryo-EM images. *Journal of Structural Biology*, 195(1):72–81, 2016.
- J. Bigot, C. Deledalle, and D. Féral. Generalized SURE for optimal shrinkage of singular values in low-rank matrix denoising. *arXiv preprint arXiv:1605.07412*, 2016.
- S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Annals of Probability*, 33(2):514–560, 2005.
- Y. Cao and Y. Xie. Low-rank matrix recovery in poisson noise. In *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, pages 384–388. IEEE, 2014.
- M. Collins, S. Dasgupta, and R. Schapire. A generalization of principal component analysis to the exponential family. *Advances in Neural Information Processing Systems (NIPS)*, 2001.
- E. Dobriban. Efficient computation of limit spectra of sample covariance matrices. *Random Matrices: Theory and Applications*, 04(04):1550019, 2015.
- E. Dobriban. Sharp detection in PCA under correlations: all eigenvalues matter. *arXiv preprint arXiv:1602.06896*, to appear in *The Annals of Statistics*, 2016.
- D. L. Donoho. Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. In *In Proceedings of Symposia in Applied Mathematics*, 1993.
- D. Donoho, M. Gavish, and I. Johnstone. Optimal shrinkage of eigenvalues in the Spiked Covariance Model. *arXiv preprint arXiv:1311.0851*, 0906812:1–35, 2013.
- M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer Science & Business Media, 2010.
- V. Favre-Nicolin, J. Baruchel, H. Renevier, J. Eymery, and A. Borbély. XTOP: high-resolution X-ray diffraction and imaging. *Journal of Applied Crystallography*, 48(3):620–620, 2015.
- T. Furnival, R. K. Leary, and P. A. Midgley. Denoising time-resolved microscopy image sequences with singular value thresholding. *Ultramicroscopy*, 2016. ISSN 0304-3991.
- K. J. Gaffney and H. N. Chapman. Imaging atomic structure and dynamics with ultrafast x-ray scattering. *Science*, 316(5830):1444–1448, 2007. ISSN 0036-8075. doi: 10.1126/science.1135923.
- M. Gavish and D. L. Donoho. The optimal hard threshold for singular values is. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.
- J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.
- D. Grün, L. Kester, and A. van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nat Methods*, 11(6):637–40, 2014.
- M. F. Hantke, T. Ekeberg, and F. R. N. C. Maia. Condor: A simulation tool for flash x-ray imaging. *Journal of Applied Crystallography*, 49(4):1356–1362, 2016.
- I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327, 2001.
- I. M. Johnstone. High dimensional statistical inference and random matrices. In *International Congress of Mathematicians. Vol. I*, pages 307–333. Eur. Math. Soc., Zürich, 2007.
- I. Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2002.
- Z. Kam. The reconstruction of structure from electron micrographs of randomly oriented particles. *Journal of Theoretical Biology*, 82(1):15–39, 1980.
- E. Katsevich, A. Katsevich, and A. Singer. Covariance matrix estimation for the cryo-EM heterogeneity problem. *SIAM Journal on Imaging Sciences*, 8(1):126–185, 2015.
- S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*, volume 3. Prentice Hall, 1993.
- O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

- S. Lee, F. Zou, and F. A. Wright. Convergence and prediction of principal component scores in high-dimensional settings. *Annals of Statistics*, 38(6):3605–3629, 2010.
- E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2005.
- J. Li and D. Tao. Simple exponential family PCA. In *AISTATS*, pages 453–460, 2010.
- J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *science*, 319(5866):1100–1104, 2008.
- N.-T. D. Loh and V. Elser. Reconstruction algorithm for single-particle diffraction imaging experiments. *Phys. Rev. E*, 80:026705, Aug 2009.
- F. Luisier, T. Blu, and M. Unser. Image denoising in mixed poisson–gaussian noise. *IEEE Transactions on Image Processing*, 20(3):696–708, 2011.
- D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.
- E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.
- F. R. Maia and J. Hajdu. The trickle before the torrent—diffraction data from X-ray lasers. *Scientific Data*, 3, 2016.
- S. Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.
- V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.*, 114(4): 507–536, 1967.
- R. J. Muirhead. *Aspects of multivariate statistical theory*. John Wiley & Sons, 2009.
- R. R. Nadakuditi. Optshrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Transactions on Information Theory*, 60(5):3002–3018, 2014.
- R. D. Nowak and R. G. Baraniuk. Wavelet-domain filtering for photon imaging systems. *IEEE Transactions on Image Processing*, 8(5):666–678, 1999.
- L. Østergaard, R. M. Weisskoff, D. A. Chesler, C. Gyldensted, and B. R. Rosen. High resolution measurement of cerebral blood flow using intravascular tracer bolus passages. part I: Mathematical approach and statistical analysis. *Magnetic Resonance in Medicine*, 36(5):715–725, 1996.
- N. Patterson, A. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190, 2006.
- D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.
- D. Paul and A. Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.
- G. Q. Qian, G. Gabor, and R. Gupta. Principal components selection by the criterion of the minimum mean difference of complexity. *Journal of Multivariate Analysis*, 49(1):55–75, 1994.
- S. Sajama and A. Orlitsky. Semi-parametric exponential family PCA. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2004.
- J. Salmon, Z. Harmany, C.-A. Deledalle, and R. Willett. Poisson noise reduction with non-local PCA. *Journal of Mathematical Imaging and Vision*, 48(2):279–294, 2014.
- P. Schwander, D. Giannakis, C. H. Yoon, and A. Ourmazd. The symmetries of image formation by scattering. II. Applications. *Opt. Express*, 20(12):12827–12849, Jun 2012. doi: 10.1364/OE.20.012827.
- S. R. Searle, G. Casella, and C. E. McCulloch. *Variance components*. John Wiley & Sons, 2009.
- A. A. Shabalin and A. B. Nobel. Reconstruction of a low-rank matrix in the presence of gaussian noise. *Journal of Multivariate Analysis*, 118:67–76, 2013.
- J.-L. Starck, F. Murtagh, and J. M. Fadili. *Sparse image and signal processing: wavelets, curvelets, morphological diversity*. Cambridge university press, 2010.
- O. Stegle, S. A. Teichmann, and J. C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.
- C. Stein. Some problems in multivariate analysis. *Technical Report, Department of Statistics, Stanford University*, 1956.
- J. A. Tropp. The Expected Norm of a Sum of Independent Random Matrices: An Elementary Approach. In C. Houdre, D. M. Mason, P. Reynaud-Bouret, and J. Rosinski., editors, *High-Dimensional Probability VII: The Cargese Volume*, Progress in Probability 71, pages 173–202. Birkhaeuser, Basel, 2016.
- R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2011.

- P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.
- K. J. Worsley, J.-I. Chen, J. Lerch, and A. C. Evans. Comparing functional connectivity via thresholding correlations and singular value decomposition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1457):913–920, 2005.
- J. Yao, Z. Bai, and S. Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, 2015.
- Z. Zhao, Y. Shkolnisky, and A. Singer. Fast steerable principal component analysis. *IEEE Transactions on Computational Imaging*, 2(1):1–12, 2016.

A Appendix

A.1 Proof of Theorem 3.1

Let $\mu = \mathbb{E}Y = \mathbb{E}A'(\theta)$ and $B_0 = \mathbb{E}YY^\top = \text{Cov}[Y] + \mu\mu^\top = \Sigma_x + \text{diag}[\mathbb{E}A''(\theta)] + \mu\mu^\top$. Let $\|\cdot\|_a$ denote a generic matrix norm, such as the operator norm or the Frobenius norm. By the triangle inequality and the Cauchy-Schwarz inequality

$$\begin{aligned} \mathbb{E}[\|S_d - \Sigma_x\|_a] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Y_i Y_i^\top - \bar{Y}\bar{Y}^\top - \text{diag}[V(\bar{Y})] - \Sigma_x\right\|_a\right] \\ &\leq \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Y_i Y_i^\top - B_0\right\|_a\right] + \mathbb{E}[\|\bar{Y}\bar{Y}^\top - \mu\mu^\top\|_a] + \mathbb{E}[\|\text{diag}[V(\bar{Y})] - \text{diag}[\mathbb{E}A''(\theta)]\|_a] \\ &\leq \left[\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n Y_i Y_i^\top - B_0\right\|_a^2\right]^{1/2} + \mathbb{E}[\|\bar{Y}\bar{Y}^\top - \mu\mu^\top\|_a] + \mathbb{E}[\|\text{diag}[V(\bar{Y})] - \text{diag}[\mathbb{E}A''(\theta)]\|_a] \end{aligned}$$

We now consider the Frobenius and operator norms separately. For the Frobenius norm, using

$$\mathbb{E}[\|\text{diag}[V(\bar{Y})] - \text{diag}[\mathbb{E}A''(\theta)]\|_{\text{Fr}}] = \mathbb{E}[\|V(\bar{Y}) - \mathbb{E}A''(\theta)\|]$$

and Propositions A.3, A.4, and A.6, we find

$$\mathbb{E}[\|S_d - \Sigma_x\|_{\text{Fr}}] \lesssim \frac{p}{\sqrt{n}}m_4 + \frac{p}{n} + \frac{\|\mu\|\sqrt{p}}{\sqrt{n}} + \frac{\sqrt{p}}{\sqrt{n}}.$$

Now, given that $m_4 = \max_i \mathbb{E}Y(i)^4$ is at least $O(1)$, the second and the last term is of smaller order than the first one. This leads to the bound

$$\mathbb{E}[\|S_d - \Sigma_x\|_{\text{Fr}}] \lesssim \sqrt{\frac{p}{n}}[\sqrt{p} \cdot m_4 + \|\mu\|].$$

For the operator norm, using

$$\mathbb{E}[\|\text{diag}[V(\bar{Y})] - \text{diag}[\mathbb{E}A''(\theta)]\|] \leq \mathbb{E}[\|V(\bar{Y}) - \mathbb{E}A''(\theta)\|]$$

and Propositions A.3, A.4, and A.7, we find

$$\mathbb{E}[\|S_d - \Sigma_x\|] \lesssim \sqrt{C(p)} \frac{(\mathbb{E}\|Y\|^4)^{1/2} + (\log n)^3 (\log p)^2}{\sqrt{n}} + \frac{p}{n} + \frac{\|\mu\|\sqrt{p}}{\sqrt{n}} + \frac{\sqrt{p}}{\sqrt{n}}.$$

This finishes the proof.

A.1.1 Sup-exponential properties

In this section we establish the sub-exponential property of our random variables. This is needed in the next sections in proving the rates of convergence.

Proposition A.1. A random variable $Y \sim p_\theta(y)$ from the exponential family is sub-exponential.

Proof. The moment generating function of Y is $\mathbb{E}[\exp(tY)] = \exp(A(\theta + t) - A(\theta))$. Since A is differentiable on an open neighborhood of θ , clearly $\mathbb{E}[\exp(tY)] \leq e$ for small t . Therefore, by the moment generating function characterization of sub-exponential random variables given in (5.16) of [Vershynin \(2011\)](#), Y is sub-exponential. \square

In the following proposition, we allow that the prior parameter θ is random, while requiring that it is bounded.

Proposition A.2. Let $Y \sim p_\theta(y)$. If θ is random and supported on a compact interval, then Y is sub-exponential.

Proof. By the characterization of sub-exponential random variables in (5.16) of [Vershynin \(2011\)](#), it is enough to show that $\mathbb{E}[\exp(A(\theta + t) - A(\theta))] \leq e$ for small t . Suppose θ is supported on $[a, b]$. Since $A(\theta + t)$ is continuously differentiable in a neighborhood of θ , we have $|A(\theta + t) - A(\theta)| \leq Ct|A'(\theta)| \leq Ct \sup_{\theta \in [a, b]} |A'(\theta)|$ for some $C > 0$, and for all t . Hence

$$\mathbb{E}[\exp(A(\theta + t) - A(\theta))] \leq \mathbb{E}[\exp(tC \sup_{\theta \in [a, b]} |A'(\theta)|)] \leq e.$$

The last inequality holds for sufficiently small t . \square

A.1.2 Auxiliary rates

Using the sub-exponential properties, we now prove the rates of convergence needed in the proof of Thm. 3.1 presented in Sec. A.1. Let $K(i) = \sup_{q \geq 1} q^{-1} (\mathbb{E}Y(i)^q)^{1/q}$ be the sub-exponential norm of the i -th coordinate of Y (see e.g., [Vershynin, 2011](#), Sec 5.2.4). By assumption, these norms are uniformly bounded, so that $K(i) \leq K < \infty$ for some universal constant K .

Proposition A.3. $\mathbb{E}[\|V(\bar{Y}) - \mathbb{E}A''(\theta)\|] \lesssim \frac{\sqrt{d}}{\sqrt{n}}$ up to universal constant factors.

Proof. By the Cauchy-Schwarz inequality, $[\mathbb{E}\|V(\bar{Y}) - \mathbb{E}A''(\theta)\|]^2 \leq \mathbb{E}[\|V(\bar{Y}) - \mathbb{E}A''(\theta)\|^2]$. Since the latter quantity decomposes into d mean squared error terms, it is enough to show that each of them is bounded by C/n up to universal constant factors. Now,

$$\mathbb{E}[V(\bar{Y}(i)) - \mathbb{E}A''(\theta(i))]^2 \leq 2\mathbb{E}[V(\bar{Y}(i)) - V(\mathbb{E}\bar{Y}(i))]^2 + 2\mathbb{E}[V(\mathbb{E}\bar{Y}(i)) - \mathbb{E}A''(\theta(i))]^2.$$

For the first term, by the Lipschitz property of V , and by the definition of K , we have

$$\mathbb{E}[V(\bar{Y}(i)) - V(\mathbb{E}\bar{Y}(i))]^2 \leq L^2 \mathbb{E}[\bar{Y}(i) - \mathbb{E}\bar{Y}(i)]^2 = n^{-1} L^2 \text{Var}Y(i) \leq n^{-1} L^2 \mathbb{E}Y(i)^2 \leq n^{-1} cL^2 K^2.$$

For the second term, notice that $A''(\theta(i)) = V(\mathbb{E}[\bar{Y}(i)|\theta(i)])$. Denoting for convenience $Z = \bar{Y}(i)$, $\alpha = \theta(i)$, this reads $A''(\alpha) = V(\mathbb{E}[Z|\alpha])$, and thus

$$T := V(\mathbb{E}\alpha) - \mathbb{E}V(\mathbb{E}[Z|\alpha]) = \mathbb{E}\{V(\mathbb{E}\alpha) - V(\mathbb{E}[Z|\alpha])\}.$$

Hence, by the Cauchy-Schwarz inequality and by the Lipschitz property of V ,

$$\mathbb{E}T^2 \leq \mathbb{E}\{V(\mathbb{E}\alpha) - V(\mathbb{E}[Z|\alpha])\}^2 \leq L^2 \mathbb{E}(\mathbb{E}\alpha - \mathbb{E}[Z|\alpha])^2.$$

Finally, the term $\mathbb{E}(\mathbb{E}\alpha - \mathbb{E}[Z|\alpha])^2 = \text{Var}(\bar{Y}(i)|\theta(i)) = n^{-1} \text{Var}(Y(i)|\theta(i)) \leq n^{-1} cL^2 K^2$ since $Y(i)$ is sub-exponential with norm at most K . Putting together all bounds, we obtain $\mathbb{E}[V(\bar{Y}(i)) - \mathbb{E}A''(\theta(i))]^2 \lesssim n^{-1}$ up to universal constant factors. By the remark in the beginning of the argument, this finishes the proof. \square

Proposition A.4. $\mathbb{E}[\|\mu\mu^\top - \bar{Y}\bar{Y}^\top\|_a] \lesssim \frac{p}{n} + \frac{\|\mu\|\sqrt{p}}{\sqrt{n}}$ up to universal constant factors, where $\|\cdot\|_a$ denotes the Frobenius norm or the operator norm.

Proof. Clearly $\|ab^\top\|_a = \|a\|\|b\|$. Then

$$\begin{aligned} \|aa^\top - bb^\top\|_a &= \|-a(b-a)^\top - (b-a)a^\top - (b-a)(b-a)^\top\|_a \\ &\leq \|(b-a)(b-a)^\top\|_a + \|a(b-a)^\top\|_a + \|(b-a)a^\top\|_a \text{ by the triangle inequality} \\ &= \|b-a\|^2 + 2\|a\|\|b-a\|. \end{aligned}$$

Using this,

$$\begin{aligned} \mathbb{E}[\|\mu\mu^\top - \bar{Y}\bar{Y}^\top\|_a] &\leq \mathbb{E}[\|\mu - \bar{Y}\|^2] + 2\mathbb{E}[\|\mu\|\|\mu - \bar{Y}\|] \\ &\lesssim \frac{p}{n} + \frac{\|\mu\|\sqrt{p}}{\sqrt{n}} \text{ by Proposition A.5.} \end{aligned}$$

□

Proposition A.5. We have $\mathbb{E}[\|\bar{Y} - \mu\|^2] \lesssim \frac{p}{n}$ and $\mathbb{E}[\|\bar{Y} - \mu\|] \lesssim \frac{\sqrt{p}}{\sqrt{n}}$ up to universal constant factors.

Proof. By the Cauchy-Schwarz inequality, $\mathbb{E}[\|\bar{Y} - \mu\|^2] \leq \mathbb{E}[\|\bar{Y} - \mu\|] \mathbb{E}[\|\bar{Y} - \mu\|]$. Then by the definition of the subexponential norm K , we have $\mathbb{E}[Y(i) - \mathbb{E}Y(i)]^2 \leq \mathbb{E}[Y(i)]^2 \leq cK^2$. Hence

$$\mathbb{E}[\|\bar{Y} - \mu\|^2] = \frac{1}{n} \sum_{i=1}^p \mathbb{E}(Y(i) - \mathbb{E}Y(i))^2 \leq \frac{cpK^2}{n}.$$

This finishes the proof. □

Proposition A.6 (Bounding the deviation of the second moment estimator for Y : Frobenius norm). Let $T_i = \frac{1}{n} (Y_i Y_i^\top - B)$ and $V_n = \sum_{i=1}^n T_i$. Then $\mathbb{E}[\|V_n\|_{\text{Fr}}^2] \lesssim \frac{p^2}{n} m_4$.

Proof. Since the Y_i are independent and identically distributed, and $\mathbb{E}T_i = 0$, we have

$$\mathbb{E}[\|V_n\|_{\text{Fr}}^2] = \mathbb{E}\left[\left\|\sum_{i=1}^n T_i\right\|_{\text{Fr}}^2\right] = n\mathbb{E}\|T_1\|_{\text{Fr}}^2 = \frac{1}{n}\mathbb{E}(\|Y_1 Y_1^\top\|_{\text{Fr}}^2 + \|B\|_{\text{Fr}}^2 - 2\text{Tr}(Y_1 Y_1^\top B)) = \frac{1}{n}(\mathbb{E}(\|Y_1\|^2)^2 - \text{Tr}(B^2)).$$

Now we can bound $\mathbb{E}(\|Y_1\|^2)^2 \leq p^2 \max_i \mathbb{E}Y_1(i)^4 \lesssim p^2 m_4$, proving the desired claim. □

Proposition A.7 (Bounding the deviation of the second moment estimator for Y : Operator norm). Let $T_i = \frac{1}{n} (Y_i Y_i^\top - B)$ and $V_n = \sum_{i=1}^n T_i$. Then

$$\begin{aligned} \mathbb{E}[\|V_n\|^2]^{1/2} &\leq \sqrt{C(p)} \|\mathbb{E}[V_n^2]\|^{1/2} + \sqrt{C(p)} \cdot \left(\mathbb{E}\left[\max_i \|T_i\|^2\right]\right)^{1/2} \\ &\lesssim \sqrt{C(p)} \frac{(\mathbb{E}\|Y\|^4)^{1/2} + (\log n)^3 (\log p)^2}{\sqrt{n}}. \end{aligned}$$

Proof. The first inequality follows directly from Theorem 5.1 in Tropp (2016). Now we find an explicit expression for the right hand side. For the first term, since the Y_i are independent and identically distributed, and the T_i are centered,

$$\mathbb{E}V_n^2 = \mathbb{E}\left(\sum_{i=1}^n T_i\right)^2 = n\mathbb{E}T_1^2 = \frac{1}{n}(\mathbb{E}\|Y_1\|^2 Y_1 Y_1^\top - B^2).$$

Since $\mathbb{E}V_n^2$ and B^2 are positive semi-definite, so $\|\mathbb{E}\|Y_1\|^2 Y_1 Y_1^\top - B^2\| \leq \|\mathbb{E}\|Y_1\|^2 Y_1 Y_1^\top\|$, we have

$$\|\mathbb{E}[V_n^2]\| \leq \frac{1}{n} \|\mathbb{E}(\|Y_1\|^2 Y_1 Y_1^\top)\|.$$

Now for any fixed vector u with $\|u\| = 1$, $u^\top \mathbb{E}(\|Y_1\|^2 Y_1 Y_1^\top) u = \mathbb{E}\|Y_1\|^2 (u^\top Y_1)^2 \leq \mathbb{E}(\|Y_1\|^2)^2$. This gives the first term, $\mathbb{E}\|Y\|^4$.

For the second term, by the triangle inequality and $(a+b)^2 \leq 2(a^2 + b^2)$,

$$\mathbb{E} \left[\max_i \|T_i\|^2 \right] = \frac{1}{n} \mathbb{E} \left[\max_i \|Y_i Y_i^T - A\|^2 \right] \leq \frac{2}{n} (\mathbb{E} \max_i \|Y_i Y_i^T\|^2 + \|B\|^2).$$

When taking square roots as required by the theorem statement, the second term in this inequality can be bounded by $\|B\| \leq \text{Tr}(B) = \mathbb{E}\|Y\|^2 \leq (\mathbb{E}\|Y\|^4)^{1/2}$. For the first term in the bound, defining $Q_i = \sum_{j=1}^d Y_i(j)^2$, we have $\|Y_i Y_i^T\|^2 = Q_i^2$, so

$$\begin{aligned} \mathbb{E} \left[(\max_i Q_i)^2 \right] &\leq \mathbb{E} \left[\left(\sum_{i=1}^n Q_i^p \right)^{2/p} \right] \text{ for } p \geq 2 \\ &\leq \left(\sum_{i=1}^n \mathbb{E}[Q_i^p] \right)^{2/p} \text{ by Jensen's inequality} \\ &\leq (n \mathbb{E}[Q_1^p])^{2/p} \\ &\leq e^2 \left((\mathbb{E}[Q_1^p])^{1/p} \right)^2 \text{ by choosing } p = \log n \\ &\lesssim [\mathbb{E}Q_1 + (\log n)^3 (\log p)^2]^2 \text{ by Lemma A.8} \\ &\lesssim [\text{Tr}(B) + (\log n)^3 (\log p)^2]^2. \end{aligned}$$

Finally, we use $\text{Tr}(B) = \mathbb{E}\|Y\|^2 \leq (\mathbb{E}\|Y\|^4)^{1/2}$ again. Putting these together leads to the result. \square

Lemma A.8. Let $Y(1), \dots, Y(p)$ be independent random variables distributed according to an exponential family $Y(j) \sim p_{\theta(j)}$ for deterministic $\theta(j) \in \mathbb{R}$. Let $Q = \sum_{j=1}^d Y(j)^2$. Define $\kappa := \frac{\sqrt{e}}{2(\sqrt{e}-1)} < 1.27$ and let η be any value in $(0, 1)$. Then for any $p \geq 1$

$$(\mathbb{E}[Q^p])^{1/p} \leq (1 + \eta) \mathbb{E}[Q] + C \frac{\kappa}{2} (1 + 1/\eta) K p^3 (\log p)^2$$

where C is small constant.

Proof. By Theorem 8 in [Boucheron et al. \(2005\)](#), we get the following Rosenthal-type bound:

$$(\mathbb{E}[Q^p])^{1/p} \leq (1 + \eta) \mathbb{E}[Q] + \frac{\kappa}{2} p (1 + 1/\eta) \left(\mathbb{E} \left[\left(\max_{j \leq d} (Y(j)^2) \right)^p \right] \right)^{1/p}$$

We proceed to bound the second term on the right hand side:

$$\begin{aligned} \left(\mathbb{E} \left[\left(\max_{j \leq d} (Y(j)^2) \right)^p \right] \right)^{1/p} &= \left(\mathbb{E} \left[\max_{j \leq d} Y(j)^{2p} \right] \right)^{1/p} \\ &\leq \left(\mathbb{E} \left[\left(\sum_{j \leq d} Y(j)^{2p \log p} \right)^{1/\log p} \right] \right)^{1/p} \end{aligned}$$

$$\begin{aligned}
&\leq \left(\left(\sum_{j \leq d} \mathbb{E} [Y(j)^{2p \log p}] \right)^{1/\log p} \right)^{1/p} \quad \text{by Jensen's inequality} \\
&\leq d^{1/p \log p} \left(\max_{j \leq d} \mathbb{E} [Y(j)^{2p \log p}] \right)^{1/p \log p} \\
&\lesssim K(p \log p)^2
\end{aligned}$$

On the last line, we have used the moments characterization of sub-exponentiality. \square

A.1.3 Proof of Theorem 4.2

It is enough to study the singular values of $\mathcal{Y}_w = n^{-1/2} \mathcal{Y}_c D_n^{-1/2}$, where $\mathcal{Y}_c = \mathcal{Y} - \bar{\mathbf{1}} \bar{\mathbf{Y}}^\top$ is the centered data matrix, because $S_w + I_p = \mathcal{Y}_w^\top \mathcal{Y}_w$. Our strategy to show that \mathcal{Y}_w is well approximated by the noise matrix $n^{-1/2} \mathcal{E}$, which is more convenient to study directly.

Indeed, since $n^{-1/2} \mathcal{E}$ has independent entries of mean 0, variance $1/n$, and fourth moment of order $1/n^2$, the distribution of the squares of its singular values converges almost surely to the standard Marchenko-Pastur distribution with aspect ratio γ . Moreover, its operator norm converges to $1 + \gamma^{1/2}$ a.s. (Bai and Silverstein, 2009).

This implies that the same two properties hold for the auxiliary matrix $\mathcal{Y}_a = n^{-1/2} (\mathcal{Y} - \bar{\mathbf{1}} A'(\theta)^\top) D_n^{-1/2}$. Indeed, we can bound the operator norm of the difference $E = n^{-1/2} \mathcal{E} - \mathcal{Y}_a$ as

$$\|E\| = \|n^{-1/2} \mathcal{E} (I_p - \text{diag}[A''(\theta)]^{1/2} D_n^{-1/2})\| \leq \|n^{-1/2} \mathcal{E}\| \|I_p - \text{diag}[A''(\theta)]^{1/2} D_n^{-1/2}\|.$$

Now $\|n^{-1/2} \mathcal{E}\| \rightarrow 1 + \gamma^{1/2}$ a.s., and $\|I_p - \text{diag}[A''(\theta)]^{1/2} D_n^{-1/2}\| \rightarrow 0$ a.s. by Lemma A.9 presented below. This shows that the spectral distribution and operator norm of \mathcal{Y}_a converge as required.

Finally, the difference of the whitened data matrix and the auxiliary matrix has rank one:

$$\mathcal{Y}_w - \mathcal{Y}_a = n^{-1/2} \bar{\mathbf{1}} (A'(\theta)^\top - \bar{\mathbf{Y}}^\top) D_n^{-1/2}.$$

Therefore \mathcal{Y}_w has the same Marchenko-Pastur limiting spectrum as \mathcal{Y}_a . This finishes the proof of Theorem 4.2.

Lemma A.9 (Convergence of empirical whitening matrix). We have $\|I_p - \text{diag}[A''(\theta)]^{1/2} D_n^{-1/2}\| \rightarrow 0$ a.s.

Proof. Since $|1 - x^{1/2}| \leq |1 - x|$ for all $x \geq 0$, it is enough to show that

$$\max_i \left| 1 - \frac{A''(\theta(i))}{V(\bar{\mathbf{Y}}(i))} \right| = \max_i \left| \frac{A''(\theta(i))}{V(\bar{\mathbf{Y}}(i))} \right| \left| 1 - \frac{V(\bar{\mathbf{Y}}(i))}{A''(\theta(i))} \right| \rightarrow 0.$$

From the expression on the right, we see that it is enough to show that $\max_i |1 - V(\bar{\mathbf{Y}}(i))/A''(\theta(i))| \rightarrow 0$ almost surely. To use the Borel-Cantelli lemma, we show how to bound the probability of $V(\bar{\mathbf{Y}}(i))/A''(\theta(i)) - 1 \geq \varepsilon$; the other direction is analogous. Since we assumed V is Lipschitz continuous with a uniform Lipschitz constant L , denoting $\delta = \varepsilon/L$, it is enough to bound the probability that $\bar{\mathbf{Y}}(i) - A'(\theta(i)) \geq \delta A''(\theta(i))$. We can write

$$\begin{aligned}
\Pr \{ \bar{\mathbf{Y}}(i) - A'(\theta(i)) \geq \delta A''(\theta(i)) \} &= \Pr \left\{ \sum_{j=1}^n Y_j(i) \geq n[A'(\theta(i)) + \delta A''(\theta(i))] \right\} \\
&\leq \mathbb{E} \exp \left[t \sum_{j=1}^n Y_j(i) \right] \exp \left[-nt[A'(\theta(i)) + \delta A''(\theta(i))] \right].
\end{aligned}$$

The moment generating function of $Y_j(i)$ is $\exp[A(\theta(i) + t) - A(\theta(i))]$, so the last quantity equals

$$\exp[n\{A(\theta(i) + t) - A(\theta(i)) - tA'(\theta(i)) - t\delta A''(\theta(i))\}].$$

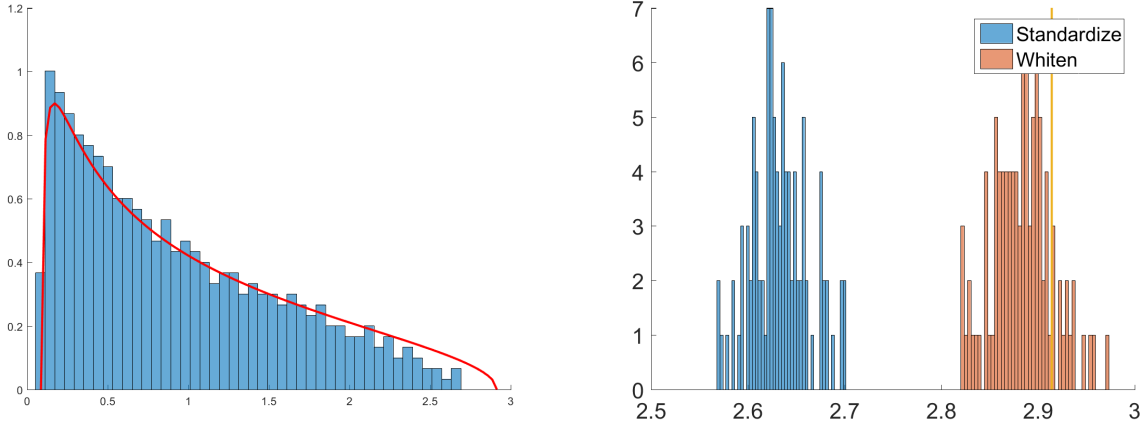


Figure A.1: Left: Histogram of one MC trial of eigenvalues after usual standardization. Right: Histogram of the upper edge of the bulk after usual standardization vs after whitening.

For t small enough (depending on A'' on a neighborhood of $\theta(i)$), this is less than $\exp[-n\delta A''(\theta(i))/2]$. Since we assumed that $A''(\theta(i)) > c$ for some universal constant $c > 0$, we get the bound $\exp[-n\delta c/2]$.

We get a similar upper bound for the probability of deviation in the other direction. We conclude that for some constants $C, c' > 0$

$$\sum_n \Pr(\max_i |1 - V(\bar{Y}(i))/A''(\theta(i))| > \varepsilon) \leq C \sum_n n \exp[-c'n] < \infty.$$

hence by the Borel-Cantelli lemma, $\max_i |1 - V(\bar{Y}(i))/A''(\theta(i))| \rightarrow 0$ almost surely. This finishes the proof. \square

A.2 Binomial observations

This section provides the details for using our methodology with binomial data. For a binomial random variable $y \sim \text{Bino}(n, q)$ with known n , the carrier measure is the discrete measure with density $\nu(dy) = \binom{n}{y}$ with respect to the counting measure on $\{0, 1, \dots, n\}$, while $\theta = \log[q/(1-q)]$, and $A(\theta) = n \log(1 + \exp(\theta))$.

Ley $Y \sim \text{Bino}(N, Q)$ be a binomial random vector with known sample sizes $N = (n(1), \dots, n(p))^\top$, and random success probabilities $Q \in \mathbb{R}^p$. After some calculation, we get $\mathbb{E}Y = \mathbb{E}(N \odot Q)$, where \odot is the elementwise product of two vectors, and

$$\text{Cov}(Y) = \text{Cov}[N \odot Q] + \mathbb{E} \text{diag}[N \odot Q \odot (\mathbf{1} - Q)].$$

The variance map for sample size n is $V(m) = m \cdot (n - m)/n$, so the estimate of the variances is $V(\bar{Y}) = \text{diag}[\bar{Y} \odot (N - \bar{Y}) \odot N]$, and the diagonally debiased estimator is $S_d = S - \text{diag}[V(\bar{Y})]$. Finally, the whitened estimator is $S_w = \text{diag}[V(\bar{Y})]^{-1/2} \cdot S_d \cdot \text{diag}[V(\bar{Y})]^{-1/2}$.

A.3 Standardization vs whitening

Following the remarks in Sec. 4.1, we present a simulation comparing the usual standardization process—dividing by the empirical standard errors of the features—to our whitening method. The results on Fig. A.1 show that in our setting usual standardization over-corrects the spectrum, so that the standard MP law is not directly a good fit. In contrast, our whitening method leads to a spectrum to which the standard MP law is a good fit.

In detail, we simulate a Poisson model where the rates are generated as $\lambda_i = m + z_i v$, where $m = (1, 1, \dots, 1)^\top$ is the all ones vector in \mathbb{R}^p , v is a vector of uniformly spaced grid points on $[-1, 1]$, and $z_i \sim \text{Unif}[-1, 1]$. We take $d = 500$ and $\gamma = 1/2$, so that $n = 1000$. Finally the data is generated as $Y_i \sim \text{Poisson}_p(\lambda_i)$. We compute the eigenvalues of the correlation matrix of the data, which is equivalent to subtracting the mean, and then standardizing each feature by its standard deviation (Fig. A.1 left). We also compute the eigenvalues of our whitened covariance matrix $S_w + I_p$ from (5) (Fig. A.1 right). In both cases we remove the top eigenvalue, which is due to the signal. We also plot a histogram of the largest noise eigenvalue after whitening as well as after standardization (Fig. A.1). On that plot, the vertical line denotes the asymptotic location of the upper edge of the Marchenko-Pastur bulk, $(1 + \gamma^{1/2})^2$.

The classical standardization leads to eigenvalues that are overall too small compared to the predictions of the MP law. This can be seen both on the plot of the histogram of all eigenvalues—where the entire distribution is shifted to the left—as well as on the plot of the histogram of the upper edge, which is also smaller than the MP upper edge.

This implies that our whitening method is more appropriate for using methods based on the MP law—such as eigenvalue shrinkage—downstream. Note that if we divide the standardized eigenvalues by their mean, we could get a good match to the standard MP law. However, this method does not seem practical, because it leads to a complicated data analytic pipeline that requires us to *guess* the signal eigenvalues, remove them, obtain an estimate of the noise level, divide the signal eigenvalues by the noise level, and then use the shrinkage methods arising from the MP law on the signal eigenvalues. For this reason, it is more convenient to use our whitening method instead.